Peer Review File

Manuscript Title: Transfer learning leveraging large-scale transcriptomics

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee expertise:

Referee #1: systems biology

Referee #2: network analysis

Referee #3: cardiac/computational biology

Referees' comments:

Referee #1 (Remarks to the Author):

The manuscript "Transfer learning leveraging large-scale transcriptomics enables predictions with limited data" by Theodoris et al introduces a machine learning algorithm of a transfer learning type, called Geneformer, which is employed on transcriptomic data with the aim of giving prediction in cases where limited samples are available. By transfer learning, the authors basically mean the following: Take one data set and learn from it. Then, learn from a different data set, with the initial conditions of the model provided from the previous learning process. This has the advantage of converging for a more accurate solution even when the number of samples is limited.
I have a few questions:

1) Genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M. While this normalization has several advantages, e.g., deprioritizing ubiquitously highly-expressed housekeeping genes and capturing important lowly expressed genes, it may lead to overestimating the value of genes that their median expression is extremely low (by normalizing them by a very low number).

2) There are several advantages of the rank-based approach. Yet, to facilitate its usage by the community, the authors may also discuss its limitations and potential caveats.

3) In silico perturbation: While the methodology of in silico deletion perturbations was demonstrated (in the Extended Methods) using fetal cardiomyocytes from the Fetal Cell Atlas, its rational or theoretical justification is not sufficiently explained. The authors may elaborate on how such perturbations can be used to reveal network connections and what's the meaning of such connections. For example, when removing a given gene from the rank value encoding of a given

single cell transcriptome, what would be the effect on the remaining genes? Does the sensitivity of the remaining genes reflect a connection between them and the removed gene?

Referee #2 (Remarks to the Author):

A. Summary of the key results

The single cell expression datasets that are available for specific tasks, e.g., for the study of rare diseases, are often very small, which limit their usability within machine learning based analytics frameworks (e.g., training a DNN often requires thousands of samples).

To benefit from the very large amount of all available single cell data when analysing smaller, task specific datasets, the authors propose Geneformer, a DNN model pre-trained on ~30 million single cell transcriptomes, which can be fine-tuned for the downstream analysis of smaller datasets.

The performance of Geneformer is assessed against ML methods trained only on the smaller datasets, showing the benefit of the transfer learning approach. In particular, Geneformer consistently boost the predictive accuracy of the downstream tasks, such as the prediction of dosage sensitivity, the prediction of chromatin dynamics, and the identification of candidate therapeutic targets for cardiomyopathy.

B. Originality and significance: if not novel, please include reference

In their paper, the authors claim that their approach proposes two main novelties:

1- From a global perspective, they propose the application of transfer learning to single cell data, leveraging the large-scale single cell data to boost the analysis of smaller, task specific datasets.

2- Within their methodology, they propose to use a novel gene ranking measure instead of directly using gene expression values. The authors claim that this non-parametric representation of gene expression will deprioritize ubiquitously highly-expressed housekeeping genes and will move to a higher rank the lowly expressed genes that can highly distinguish cell states.

On one hand, the large number of downstream analyses used to evaluate Geneformer is impressive, and shows that the method can indeed boost the predictive accuracy of many downstream tasks, with respect to analyses using only the smaller, task specific datasets.

On the other hand, the application of transfer learning for the analysis of single cell data is not new, and many ML methods/transfer learning strategies have already been proposed, e.g.:

- In [Stumpf et al., Commun Biol 3, 736 (2020)] a multiclass logistic regression model is first pre-trained on Mouse bone marrow single cell expression data to recognize different cell types, and then is fine-tuned on Human single cell expression data, showing 83% overall accuracy.

- In [Wang et al., Genome Biol 20, 165 (2019)], a deep encoder learns the low dimensional embeddings of cells from their single cell expression data, and the authors actively use transfer loss across different datasets to remove batch effect.

- In [Lieberman et al., PLoS ONE 13(10): e0205499 (2018)] the authors propose a framework for classifying cells into cell types according to their single gene expression data, centered around a pre-trained XGBoost classification model that is transferred for the classification of smaller datasets.

- …

As exemplified above, the literature is rich with such approaches, which are not cited / compared with in this paper. In their paper, the authors only mention applications of transfer learning to non-biological field, such as natural language understanding and computer vision. Because of this, it is unclear how the proposed approach differs from the existing methods and if it is any better. At best, the proposed study further confirms the potential benefit of transfer learning for analyzing single cell expression data.

Furthermore, the benefit of the novel gene ranking strategy is not evaluated.

C. Data & methodology: validity of approach, quality of data, quality of presentation.

The methodology is explained well and is sounds.
However, a key issue is that the presented approach, which is pre-trained on large datasets and then fine-tuned for a specific task on smaller datasets, is only compared to simple ML methods that are trained solely on the smaller, task-specific datasets (e.g., SVM, linear regression, or the proposed DNN, but without the pre-training on the large dataset). Importantly, the proposed approach is not compared to any other transfer learning methods, while many have been proposed in the literature for the study of single cell expression data (e.g., see point B). Thus, while the study further sheds light on the benefit of transfer learning to study single cell data, it does not allow for assessing the quality of the proposed DNN architecture / transfer learning strategy.

Furthermore, in their approach, the authors make many methodological choices, e.g., using an attention based DNN with 6 encoder units (each composed of a self-attention layer and of a feed forward layer), using ranks instead of gene expression values, using 512 dimensions for embedding the cells, using precisely 15% of randomly chosen genes in the masked learning objective for the pre-training step, etc… While the intuitions / reasoning behind some of these choices are given, the authors do not present the results showing their benefits.

As I am not familiar with the usage of the single cell expression data used in the study, I cannot comment on the filtering strategy used by the authors. From a pure ML point of view, it would be nice if the authors could report / comment on the potential overlap between the large pre-training dataset (from Geneformer-30M) and the smaller, task specific datasets used in the downstream analyses, which could lead to ‘data leakage’.

D. Appropriate use of statistics and treatment of uncertainties

Statistics are properly used.

E. Conclusions: robustness, validity, reliability

The robustness of the fine-tuning step is properly evaluated with 5-fold cross validation.

My main concern here is the robustness of the pre-training step, which is not evaluated. In particular, the pre-training process, which is the most computationally intensive step, is heuristically done using a masked learning objective, in which 15% of randomly chosen expressed genes per sample are masked and used as targets for the training process. Due to the randomness of this training process, a robustness analysis is needed.

F. Suggested improvements: experiments, data for possible revision

To better position their approach, the authors should include a review of the literature on the transfer learning approaches that have been applied to the analysis of single cell expression datasets.

Based on this review, the authors should compare the performance of their proposed Geneformer to those of relevant transfer learning approaches for single cell expression data analysis from the literature (or at least, they should explain why the comparison is not possible).

The robustness of Geneformer's pre-training step should be evaluated in a cross-fold experiment (evaluating how the ability of Geneformer to boost downstream analysis is impacted by the randomness of the masked learning strategy).

The different methodological choices should be evaluated to show their contributions; e.g., the results obtained by using the proposed gene ranking strategy should be evaluated against those that are obtained when using traditional (normalized, batch processed) gene expression values.

G. References: appropriate credit to previous work?

The authors do not mention / cite any previous work on transfer learning applied on single cell data, while many approaches have been proposed (e.g., the list in point B above).

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Given the above-mentioned problems of lack of literature survey on transfer learning applied to single cell data and the lack of comparison with such methods, the paper (including abstract, introduction, and conclusion) does not clearly / objectively present the context of the paper.

Referee #3 (Remarks to the Author):

Summary

Theodoris et al. developed a computational tool called Geneformer which utilises a compilation of 30M publicly available large-scale single-cell transcriptomes from various tissue and cell types, labelled Genecorpus-30M. Geneformer is pre-trained using a transformer encoder network on Genecorpus-30M to identify the context-specific gene regulatory and expression hierarchies. The authors demonstrate that gene-former is context-aware and is able to characterise network hierarchy through the neural network structure. With fine-tuning using limited data specific to the task, geneformer was capable of predicting known regulators of cell identity, dosage-sensitive genes, bivalency, downstream targets, hierarchy of gene networks and transcription factor cooperativity. The authors further use geneformer to model in-silico deletion to identify disease and cell-identity related genes and screen for potential druggable candidates. Given its broad utility even with limited input data, its applicability in demonstrated and many related subjects would be enormous. However, we have found major flaws and limitations of the study, mainly lack of new biological insight extracted from predictions as well as algorithmic and conceptual advances in the computational method and analysis results are not sufficient to support conclusions.

Major limitations

1. Absence of new biological insight in predictions

All predictions in the study were validated only with previously published literature (i.e. recovery of what is already known). While this is still a valid approach to demonstrate success of Geneformer as a computational tool, it would be much desirable if new predictions were followed up with experimental validation. Importantly, this means that the study did not sufficiently demonstrate discovery capacity of Geneformer. In the revision, it would be great to see if authors can follow up with a few predictions, for example in silico deletion of predicted genes influence downstream gene network. This establishes essential causal relationship between the targeted gene and its biological effect, which lacks in the current study.

2. Suboptimal comparisons for computational algorithm and output

A few limitations related to performance testing and output interpretation. First, performance analysis of Geneformer largely focused on recovery of true positives vs. false positives in ROC. ROC curves can be misleading if data are imbalanced but it is not clear how balanced (or imbalanced) the testing data. Precision-recall curve (PRC) helps address this issue especially huge imbalance is present in the data – which is possible given positives are much rarer than negatives in these contexts. This may not be necessary but at least would be ideal to declare if data imbalance exists in test sets.

Second, effect analysis of in silico deletion of GATA4 & TBX5 (Fig 5b) was compared between known co-bound targets vs housekeeping genes. This analysis is incomplete and potentially misleading as the comparison did not show its effect on other TF targets. Without knowing this, we cannot conclude that Geneformer is capable of predicting correct network hierarchy specific to a given context, hence context-aware. Is the effect broader over TFs or really specific for these targets? This

discrepancy needs to be clearly answered by for example comparing against randomly chosen TF targets.

3. Magnitude of in-silico deletion effects
Throughout the paper, in-silico deletion of genes is used to demonstrate context-awareness and model disease and/or potential drugs. Firstly, Effect analysis of in silico expression of OKSM factors (Extended Data Fig 1C) which was used to demonstrate the context-awareness of Geneformer was noted to significantly shift gene embeddings toward an iPSC state. However, a comparison point of gene-embeddings in an iPSC state should be provided to show the magnitude of the shift. Furthermore, the gene embeddings after artificially expressing OKSM could be compared to the gene embeddings from the iPSC reprogramming dataset presented in Extended Data Fig 1B. Demonstration of context-awareness could also be performed with additional reprogramming factors, for instance MyoD.

Secondly, in the in-silico therapy of cardiomyopathy, gene candidates were identified which shifted cells from a non-failing state to a cardiomyopathy state, and then druggable candidates which shifted the cells to a non-failing state. The magnitude of these shifts is unclear in the results – it would be beneficial to show how much a gene's deletion shifts cell embeddings toward the cardiomyopathy or non-failing state respectively.

Finally, if authors could calculate how faithfully in-silico deletions model in-vitro knock outs on gene embeddings. This could be performed using existing RNA datasets of gene knockouts.

4. The benefits from transforming UMI count data to rank-based matrix need to be justified and technical issues related to using the rank metric with regard to genes sharing the same rank need to be assessed. Here, scRNAseq UMI count data is transformed into rank-based data for inputting into the neural network and the authors claimed that this transformation is a novel contribution of their method, however the benefits of this rather simple preprocessing step need more evidence.

The rank transformation may cause issues because many genes may share the same rank because in a scRNAseq UMI count data, the value range is much less diverse compared to bulk sequencing. The issue of shared ranks is even more common for many genes with 0 UMI count in any given cell. This is a classical issue due to the inherent sparsity of the single cell data. The authors should assess the effect of discretising the UMI data to rank data, with regard to how the data distribution changes and how these changes affect downstream analyses.

5. More normalisation is required to provide evidence on how the self attention transformer results in the batch effect removal

In their workflow, for each gene, they use global median expression value for only non-zero value, discarding the cases where the gene has 0 value. This approach does not address the uneven distribution of zero values across genes, cells and contexts. The number of 0 values is also dependent on sequencing depth, with more 0s expected if the depth does not reach a saturation level. Normalisation sequencing depth does not solve this issue. In the extended methods, the authors described normalisation across samples based on sequencing depth (to total transcript

count per cell) followed by normalising gene counts using non-zero median expression values across all samples (although not clearly described how this is done) before ranking genes for each cell. Also the aggregation of data sub(sets) among the 30M single cells does not account for cell type information, nor sample-to-sample normalisation is used.

Consequently, the results showing the normalisation effect are not convincing. Extended Figure 1a shows results for assessing batch effect, but the density plots do not show how the embedding comparisons reflect the similarity/dissimilarity in the original values. The authors show that embeddings of gene A and gene B in the same cell are more different than those of gene A vs gene A in the same cell type regardless of the batch effects (cosine similarity). The lower similarity between two genes in the same cell relative to the same gene between two different cells is expected and is not sufficient to suggest that the embedding is context dependent. The authors described that they compared UMAP of cells based on the embeddings with UMAP of original gene expression space. However, the resulting UMAPs are different with a clear reduction in the ability to separate cells in different conditions (extended Figure 2a) . Extended Figure 2f shows some improvement in data integration, but the UMAPs still suggest suboptimal integration, because cardiomyocyte 1 and 2 are still separated by platforms

6. Interpretability of the network needs to be improved. The authors have not adequately described how the attention layers learn the gene and cell context, but rather assume the attention heads in the pretraining step can learn the gene and cell contexts across the 30M transcriptomes. The authors described the analysis of attention weights extracted from attention heads could capture gene regulation network hierarchy and show examples in Figure 4. The concept on how the attention heads capture the gene hierarchy with regards to which gene pays more attention to other genes is described, but figures to prove the model captures this hierarchy is not convincing. For example, the attention for NOTCH1 is higher in layers o, 1 but lower in subsequent layers compared to TCF4 and SOX7 (Figure 4e). This variation from layers to layers and heads to heads appear to be independent of biological context. Such random variation is also observed in Figure 4g. In addition, the interpretation of gene embedding is also questionable. In extended Figure 1 d, why are there a large subset of the 256 latent dimensions that have low embedding values across all genes all cell types?

7. Code is not available. To assess the model in this work, in terms of reproducibility and performance, it is essential to access to the source code and the scripts used to generate figures. The description of model architecture and concepts in the manuscript is not sufficient to fully prove if the model is working, and if the figures are reliable and reproducible. Overall the model architecture is relatively simple with 6 transformer units, each containing a self attention layer (4 attention heads) and a feed forward fully connected network (512 neurons) to get embeddings in 256 dimensions. For benchmaring with other methods, this work only compared to simple classifiers like random forest, support vector machine and logistic regression, but not with other deep learning methods. The codes for comparisons need to be available to assess the reliability and reproducibility.

Minor limitations
1. Practicality of the tool for prediction

It is important for readers to know practicality of this tool when applying their given input of interest. Importantly, the paper did not indicate how much input data are needed to yield confident predictions. The authors used limited data for fine-tuning (e.g. 30k for gene network hierarchy, which is certainly not small in many cases) are sufficient but it's not clear how this variable influences the outcome. Are there any thresholds (i.e. minimum number required?) or totally data-dependent? I assume there should be a certain range of cell numbers needed to adjust the pre-defined model assuming acceptable sequencing depth. This can be an important question to be answered as often not such cell numbers are available for a cell type of interest.

2. Unclear justification of the model selection
Authors defined the neural network with specified hyperparameters (e.g. 5 layers, 256 dimensions etc.) without clear justification. While these parameters were presumably selected for their optimal performance, it remains unclear to readers how this was done relevant details were not included. It would be ideal to include such details so that the model selection was done in a reproducible way. It is understandable that this particular ML model can often appear as a "blackbox" due to low-interpretability. Despite that, I think it's critical to justify selection of a given model from computational perspective – how the training was performed (e.g. separation of datasets for training, validation and optimization).

3. Genes affected by in-silico deletion
If authors could identify downstream genes which were affected by the in-silico deletions performed in Fig 2d and characterise them using curated databases such as KEGG or GO, it would help provide more context for the mechanism of the in-silico deletions and their downstream effects.

3. Normalisation method
Authors ranked genes within each cell by normalising by that gene's expression across the entire Genecorpus. Is it possible that this strategy may prioritise genes that are expressed highly in a cell-type specific manner, over genes expressed lowly, such as transcription factors?

4. In silico over-expression of genes
Aside from the OKSM factors, the paper only utilises in silico deletion. Only utilising deletion could miss important biological aspects, particularly in disease and therapeutic contexts. Can in-silico overexpression be applied to the cardiomyopathy dataset?

5. Interpretation of significance
Some of result data show dubious significance, for example in Fig 6d. These targets are statistically marginally enriched. As p-value is largely affected by the sample size (assuming n=104), would fold-change be a useful alternative?

6. Absence of scale bar
Figure 6e contour plot does not include the scale bar for gene number while the interpretation of the result is not affected by this information.

Author Rebuttals to Initial Comments:

The reviewers' comments are shown in **black** and our responses are in **blue.**

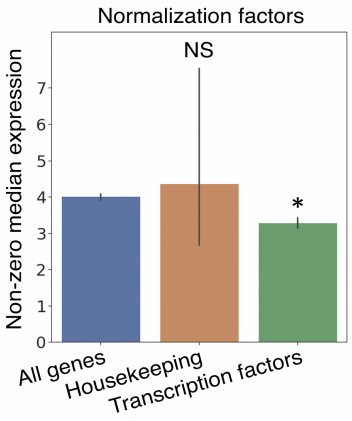
Referee One:

The manuscript "Transfer learning leveraging large-scale transcriptomics enables predictions with limited data" by Theodoris et al introduces a machine learning algorithm of a transfer learning type, called Geneformer, which is employed on transcriptomic data with the aim of giving prediction in cases where limited samples are available. By transfer learning, the authors basically mean the following: Take one data set and learn from it. Then, learn from a different data set, with the initial conditions of the model provided from the previous learning process. This has the advantage of converging for a more accurate solution even when the number of samples is limited. I have a few questions:

1) Genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M. While this normalization has several advantages, e.g., deprioritizing ubiquitously highly-expressed housekeeping genes and capturing important lowly expressed genes, it may lead to overestimating the value of genes that their median expression is extremely low (by normalizing them by a very low number).

Authors' response: We thank the reviewer for this comment. We quantified the normalization factors for housekeeping genes and transcription factors compared to all genes to better characterize the effects of the normalization (below and in Extended Data Fig. 1c). Transcription factors are normalized by a statistically significantly lower factor (resulting in higher prioritization in the rank value encoding) compared to all genes. Housekeeping genes on average show a trend of a higher normalization factor (resulting in deprioritization in the rank value encoding) compared to all genes. Thus, genes that are expressed at extremely low levels will indeed be normalized by a low number, but this is the intended effect in order to further prioritize genes like transcription factors that are important to informing the model about network dynamics but would be deprioritized in the rank value encoding if not normalized by their non-zero median value of expression.

Manuscript changes: Addition of Extended Data Fig. 1c (below):



2) There are several advantages of the rank-based approach. Yet, to facilitate its usage by the community, the authors may also discuss its limitations and potential caveats.

Authors' response: We thank the reviewer for raising this important point. To ensure a balanced discussion of our approach, we added the following commentary to the main text regarding limitations of the rank-based representation.

Manuscript changes: Addition of the following text: "Although the rank-based representation has limitations including not fully taking advantage of the precise gene expression measurements provided in transcript counts, the rank value encoding provides a nonparametric representation of each cell's transcriptome and takes advantage of the many observations of each gene's expression across Genecorpus-30M to prioritize genes that distinguish cell state."

3) In silico perturbation: While the methodology of in silico deletion perturbations was demonstrated (in the Extended Methods) using fetal cardiomyocytes from the Fetal Cell Atlas, its rational or theoretical justification is not sufficiently explained. The authors may elaborate on how such perturbations can be used to reveal network connections and what's the meaning of such connections. For example, when removing a given gene from the rank value encoding of a given single cell transcriptome, what would be the effect on the remaining genes? Does the sensitivity of the remaining genes reflect a connection between them and the removed gene?

Authors' response: We thank the reviewer for raising this point. We designed an in silico perturbation approach where the rank of given genes is perturbed to model their inhibition or activation. The effects of the in silico perturbation are measured at the cell and gene embedding level, modeling how the perturbation affects the cell's state and the expression of downstream genes within the gene network, respectively. Because the model is trained to understand which genes attend to which other genes within the network, we theorized that genes most affected by a given gene's perturbation would be enriched for that gene's biological downstream targets.

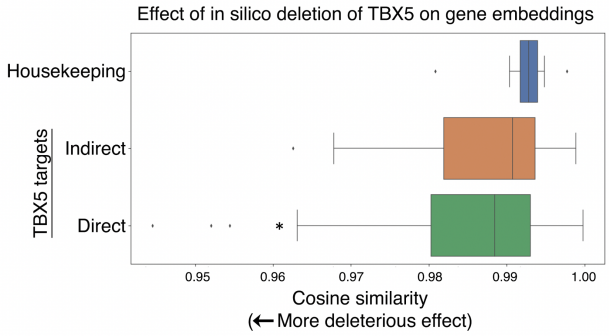
Indeed, we found that to be the case, for example when modeling GATA4 inhibition in fetal cardiomyocytes. In silico deletion of GATA4 had a significantly higher effect on genes known to be most significantly dysregulated by GATA4 variants in a previously reported iPSC disease

model of GATA4-related heart defects[1] (Extended Data Fig. 9d). Furthermore, direct GATA4 targets (as defined by GATA4 binding by ChIP-seq in the iPSC disease model[1]) were significantly more impacted by in silico deletion of GATA4 in fetal cardiomyocytes compared to housekeeping genes or indirect targets (Fig. 5a).

Similarly, direct TBX5 targets were significantly more impacted by in silico deletion of TBX5 in fetal cardiomyocytes as compared to housekeeping genes or indirect targets (Extended Data Fig. 9e). As the reviewer suggested, our aggregate data indicate that genes with increased sensitivity to a particular gene's deletion reflect genes that are biological targets within the network. We have added additional commentary (below) regarding the rationale behind the in silico perturbation experiments to the manuscript to further clarify this.

Manuscript changes:

- Addition of Extended Data Fig. 9e (below):



- Addition of the following text:

"We designed an in silico perturbation approach where the rank of given genes is perturbed to model their inhibition or activation. The effects of the in silico perturbation are measured at the cell and gene embedding level, modeling how the perturbation affects the cell's state and the expression of downstream genes within the gene network, respectively."

"Analogously, in silico deletion of TBX5, another known congenital heart disease gene, in fetal cardiomyocytes more significantly impacted its direct targets (as defined by ChIP-seq) compared to indirect targets and housekeeping genes (Extended Data Fig. 9e). These data suggest that in silico perturbation can be applied to model gene network connections."

Referee Two:

The single cell expression datasets that are available for specific tasks, e.g., for the study of rare diseases, are often very small, which limit their usability within machine learning based analytics frameworks (e.g., training a DNN often requires thousands of samples).

To benefit from the very large amount of all available single cell data when analysing smaller, task specific datasets, the authors propose Geneformer, a DNN model pre-trained on ~30 million single cell transcriptomes, which can be fine-tuned for the downstream analysis of smaller datasets.

The performance of Geneformer is assessed against ML methods trained only on the smaller datasets, showing the benefit of the transfer learning approach. In particular, Geneformer consistently boost the predictive accuracy of the downstream tasks, such as the prediction of dosage sensitivity, the prediction of chromatin dynamics, and the identification of candidate therapeutic targets for cardiomyopathy.

B. Originality and significance: if not novel, please include reference

In their paper, the authors claim that their approach proposes two main novelties:

1- From a global perspective, they propose the application of transfer learning to single cell data, leveraging the large-scale single cell data to boost the analysis of smaller, task specific datasets.

2- Within their methodology, they propose to use a novel gene ranking measure instead of directly using gene expression values. The authors claim that this non-parametric representation of gene expression will deprioritize ubiquitously highly-expressed housekeeping genes and will move to a higher rank the lowly expressed genes that can highly distinguish cell states.

On one hand, the large number of downstream analyses used to evaluate Geneformer is impressive, and shows that the method can indeed boost the predictive accuracy of many downstream tasks, with respect to analyses using only the smaller, task specific datasets.

On the other hand, the application of transfer learning for the analysis of single cell data is not new, and many ML methods/transfer learning strategies have already been proposed, e.g.:

- In [Stumpf et al., Commun Biol 3, 736 (2020)] a multiclass logistic regression model is first pre-trained on Mouse bone marrow single cell expression data to recognize different cell types, and then is fine-tuned on Human single cell expression data, showing 83% overall accuracy.

- In [Wang et al., Genome Biol 20, 165 (2019)], a deep encoder learns the low dimensional embeddings of cells from their single cell expression data, and the authors actively use transfer loss across different datasets to remove batch effect.

- In [Lieberman et al., PLoS ONE 13(10): e0205499 (2018)] the authors propose a framework for classifying cells into cell types according to their single gene expression data, centered around a pre-trained XGBoost classification model that is transferred for the classification of smaller datasets.

- …

As exemplified above, the literature is rich with such approaches, which are not cited / compared with in this paper. In their paper, the authors only mention applications of transfer learning to non-biological field, such as natural language understanding and computer vision. Because of this, it is unclear how the proposed approach differs from the existing methods and if it is any better. At best, the proposed study further confirms the potential benefit of transfer learning for analyzing single cell expression data.

Authors' response: We thank the reviewer for raising this point for clarification. There are three major distinctions between the work we present here and the prior references to "transfer learning" in the biological field.

Firstly, many prior reports of "transfer learning" in biology involve training a model on a specific task on one dataset, and then using the model to make predictions on the same task in another dataset. For example, the more recently reported scDeepSort[2] claims to use a "pretraining" approach because they train their model on a larger labeled dataset to distinguish cell types and then use this "pretrained" model to distinguish cell types in new datasets. However, there is minimal distinction between what they refer to as "pretraining" and "transfer learning" from simply "training" and "learning" as they are training the model on the same learning objective as is being used for prediction. Essentially, there is no transfer of learning from one task to another.

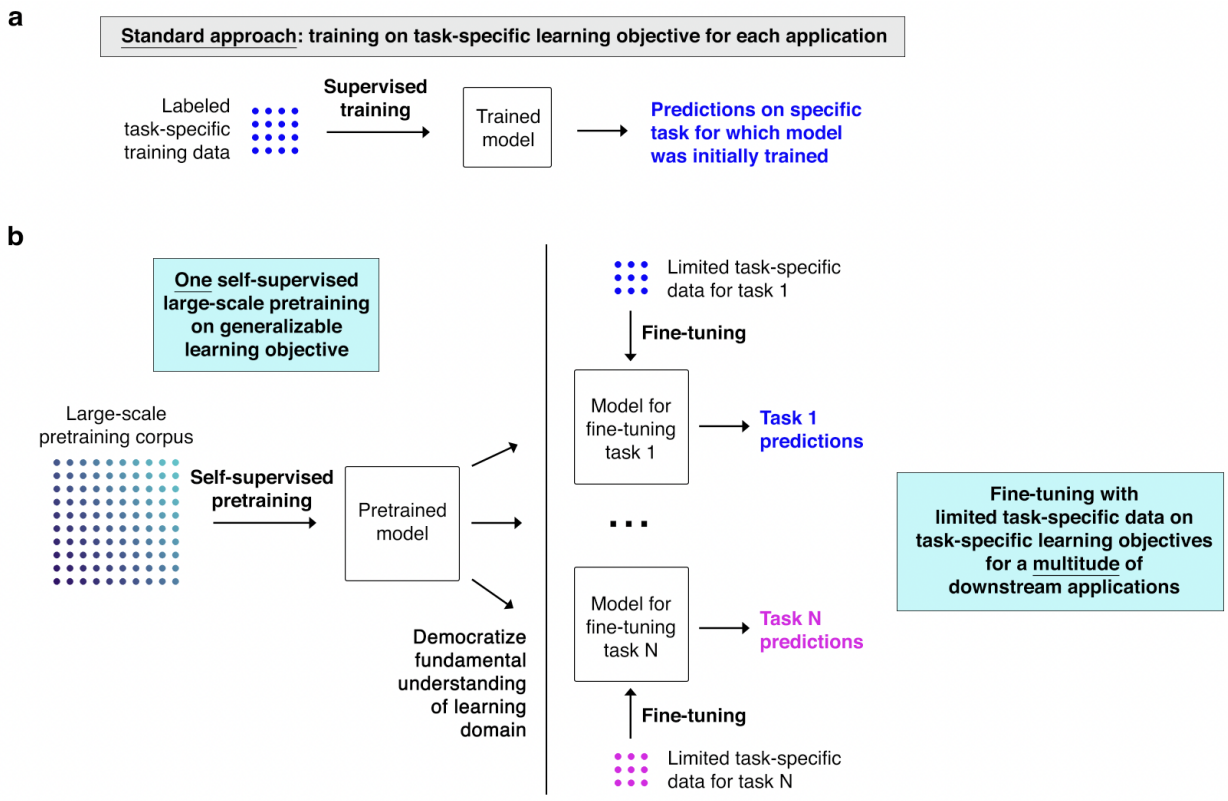
CaSTLe[3] takes a similar approach where the model is trained on cell type data to predict on the same cell type annotation task in a second dataset. In our approach and the cited references in the NLU and computer vision fields, models are pretrained on large-scale datasets with a learning objective meant to gain a fundamental understanding of the informational domain that is transferable to a multitude of downstream tasks. These pretrained models are then fine-tuned with limited task-specific data on a separate downstream learning objective, transferring previously gained knowledge to boost predictions in this downstream task despite limited task-specific data.

This leads to the second distinction, which is that our approach and the cited references in the NLU and computer vision fields democratize the fundamental knowledge learned from the large-scale pretraining to a multitude of downstream tasks. The cited references train models on

a narrow, specific task, such as cell type classification, and then apply the trained models to make predictions within that same narrow task in other datasets.

For example, in scDeepSort and CaSTLe, the authors train a new model for each separate tissue and only use that model to predict cell types in that specific tissue. This approach is impractical because it is infeasible and inefficient to retrain the model for every specific downstream task when incorporating data at the scale we have done in this present work.

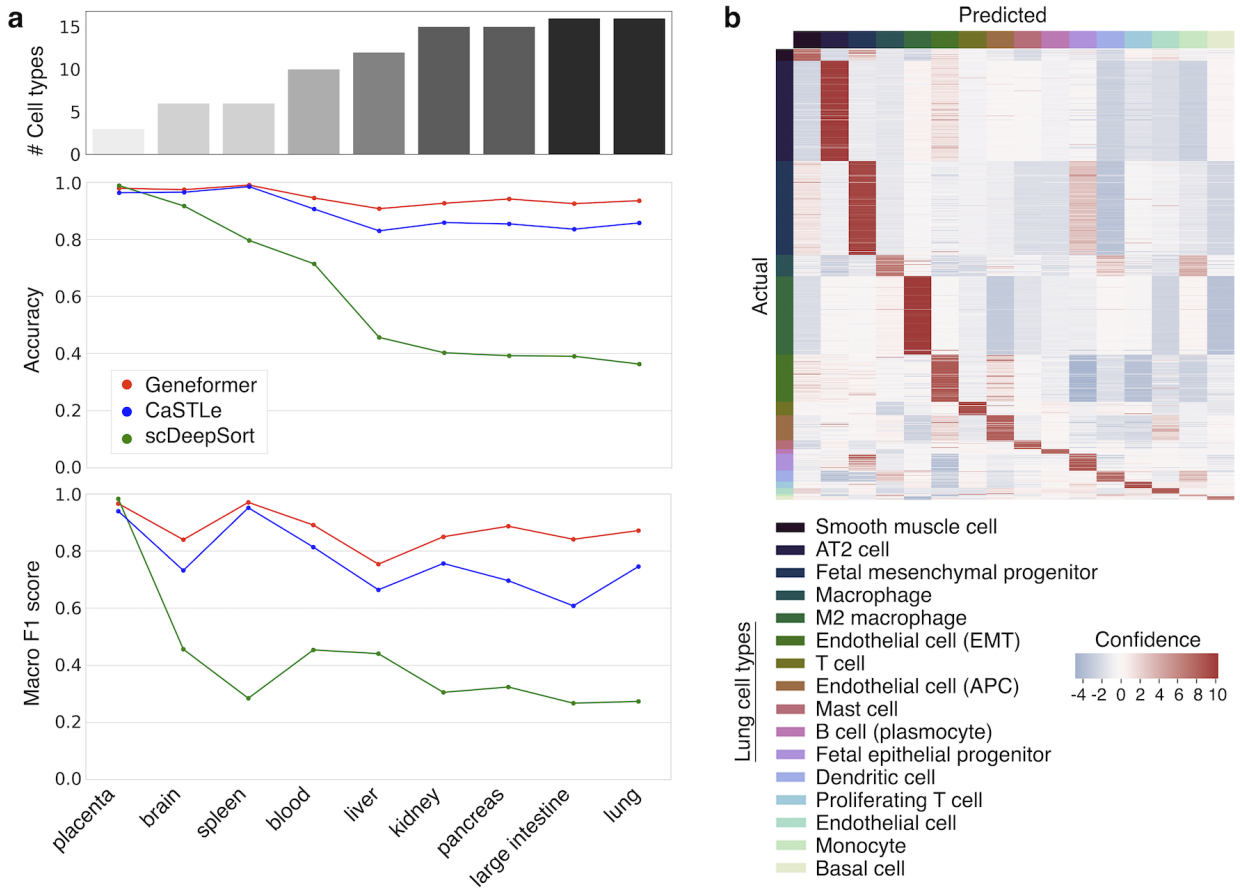
Our approach of using pretraining to imbue a model with fundamental knowledge transferrable to many downstream tasks is a major conceptual distinction between this work and the current standard in the biological field (below and Extended Data Fig. 1a-b; **a,** Standard approach, **b,** Geneformer approach).



Thirdly, another major distinction is the approach of self-supervised learning in the pretraining stage. The use of a masked learning objective in this work and other fields such as NLU allows the incorporation of large-scale data at the pretraining stage without a requirement of any training labels. Importantly, we have incorporated orders of magnitude of more pretraining data than would be possible if requiring supervised pretraining, and incorporation of larger amounts of pretraining data consistently improves predictions in downstream tasks, as demonstrated in Fig. 2b.

We do appreciate the crux of reviewer's point, namely that there is considerable ambiguity in the use of the term "transfer learning" in the existing literature. We appreciate the opportunity to illustrate the distinction between our work and the existing literature and to highlight the conceptual advances in our approach.

Finally, to be sensitive to the reviewer's concerns which we appreciate are likely to be shared by many readers, we have added benchmarking comparisons of Geneformer to CaSTLe and scDeepSort, and found that Geneformer improved predictions in a variety of human tissues by both accuracy and macro F1 score. The gap in performance increased as the number of cell type classes increased, indicating that Geneformer was robust even in increasingly complex multiclass prediction applications (below and Extended Data Fig. 5-6).



a, *Predictive potential (as measured by accuracy and macro F1 score) of Geneformer fine-tuned for cell type annotation in the indicated human tissues as compared to XGBoost (CaSTLe) and deep neural network-based (scDeepSort) methods. The top bar graph indicates the number of cell type classes for each tissue.* ***b,*** *Out of sample predictions by Geneformer fine-tuned to distinguish lung cell types (training on 80% of cells, predictions on held-out 20% of cells).*

Of note, we did not compare performance to Stumpf et al. 2020 because this model only predicts bone marrow cell types, whereas we could compare performance to CaSTLe and scDeepSort in multiple different tissues. Also, we did not compare performance to Wang et al.

2019 because batch effect removal is not a primary goal of Geneformer and would require building considerable infrastructure as a fine-tuning application to properly compare to this tool.

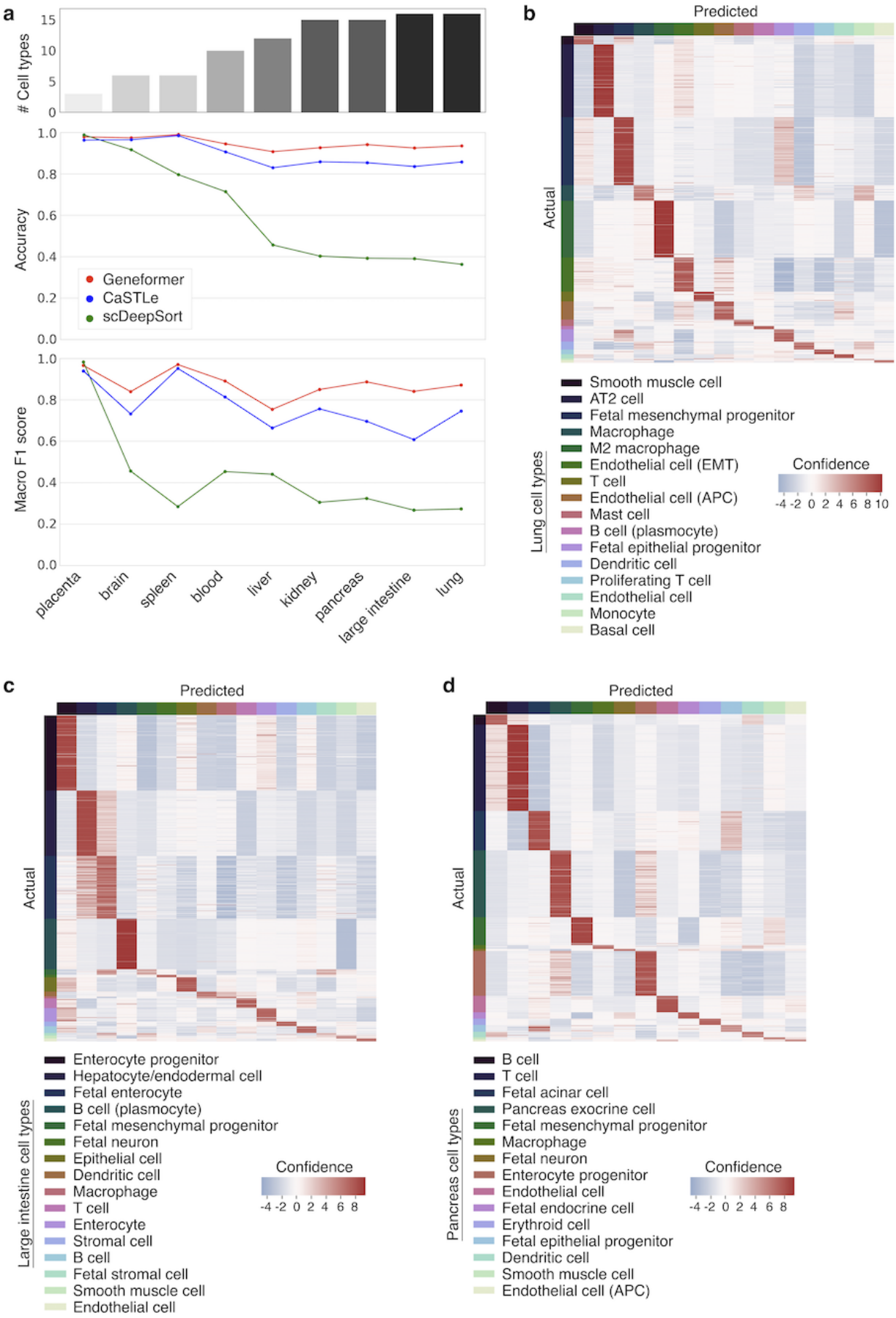
Manuscript changes:

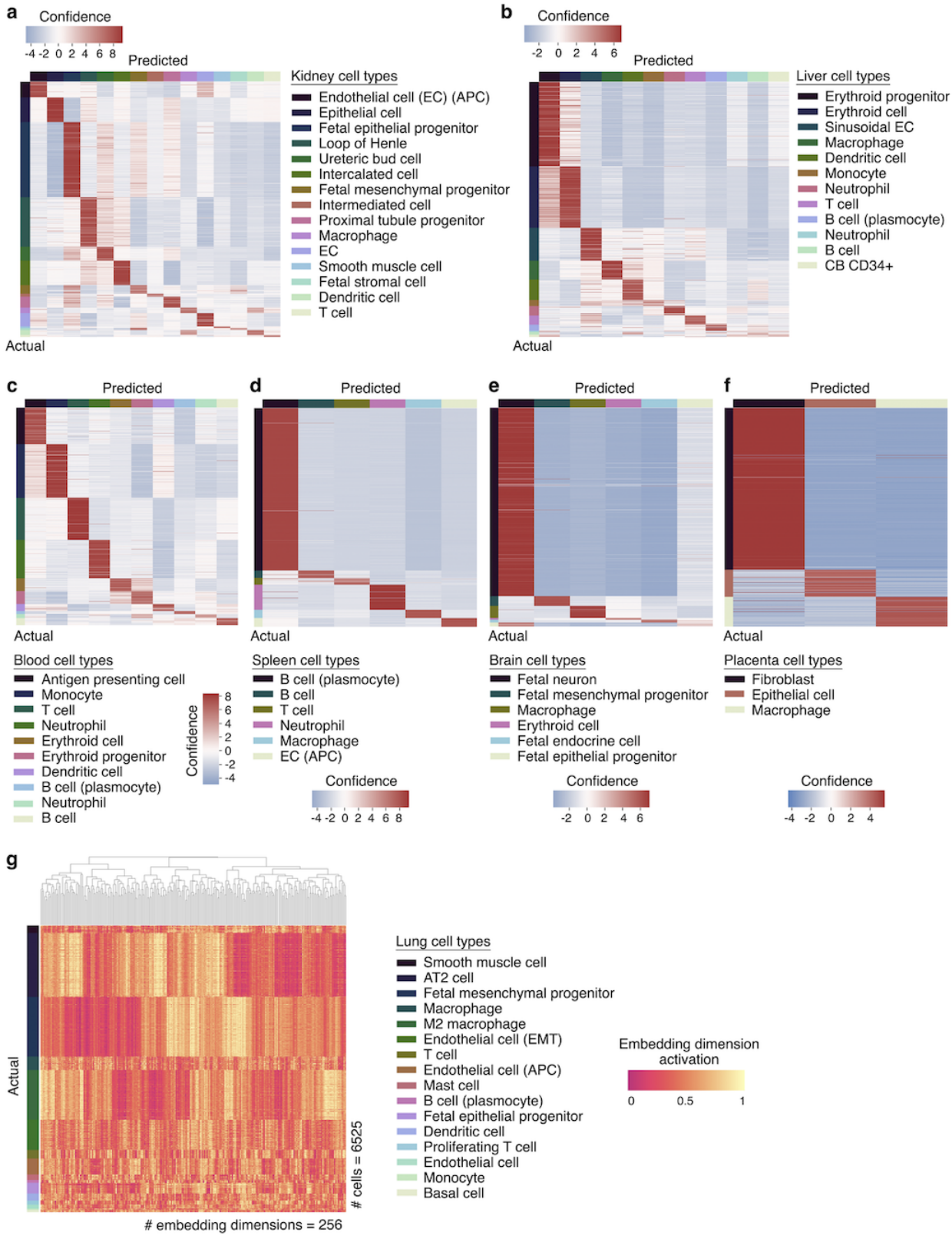
- Addition of Extended Data Fig. 1a-b (above) and relevant text:

“Unlike modeling approaches that necessitate retraining a new model from scratch for each task, this approach democratizes the fundamental knowledge learned during the large-scale pretraining phase to a multitude of downstream applications distinct from the pretraining learning objective, transferring knowledge to new tasks (Fig. 1a, Extended Data Fig. 1a-b).”

- Addition of Extended Data Fig. 5-6 (below) and relevant text:

“Although Geneformer is most focused on understanding network dynamics rather than cell-level annotations, we further investigated Geneformer’s performance in cell type annotation given it is a common application for previously published models. We compared Geneformer to alternative XGBoost and deep neural network-based models. These methods train a new model from scratch for each separate tissue using the same supervised learning objective as is used for the final cell type predictions in that specific tissue. Therefore, these approaches do not take advantage of the large amounts of data available more broadly that are not specifically labeled for that task. In contrast, Geneformer learns from large-scale unlabeled data during the self-supervised pretraining using a generalizable learning objective to gain fundamental knowledge that can then be transferred to a multitude of new and diverse fine-tuning tasks. Compared to these alternative methods, Geneformer boosted cell type predictions in a variety of tissues, with the gap in performance by accuracy and macro F1 score increasing as the number of cell type classes increased, indicating that Geneformer was robust in even increasingly complex multiclass prediction applications (Extended Data Fig. 5-6). “


a
Cell types
Accuracy
Macro F1 score
Geneformer
CaSTLe
scDeepSort
placenta
brain
spleen
blood
liver
kidney
pancreas
large intestine
lung
b
Predicted
Actual
Lung cell types
Smooth muscle cell
AT2 cell
Fetal mesenchymal progenitor
Macrophage
M2 macrophage
Endothelial cell (EMT)
T cell
Endothelial cell (APC)
Mast cell
B cell (plasmocyte)
Fetal epithelial progenitor
Dendritic cell
Proliferating T cell
Endothelial cell
Monocyte
Basal cell
Confidence
c
Predicted
Actual
Large intestine cell types
Enterocyte progenitor
Hepatocyte/endodermal cell
Fetal enterocyte
B cell (plasmocyte)
Fetal mesenchymal progenitor
Fetal neuron
Epithelial cell
Dendritic cell
Macrophage
T cell
Enterocyte
Stromal cell
B cell
Fetal stromal cell
Smooth muscle cell
Endothelial cell
Confidence
d
Predicted
Actual
Pancreas cell types
B cell
T cell
Fetal acinar cell
Pancreas exocrine cell
Fetal mesenchymal progenitor
Macrophage
Fetal neuron
Enterocyte progenitor
Endothelial cell
Fetal endocrine cell
Erythroid cell
Fetal epithelial progenitor
Dendritic cell
Smooth muscle cell
Endothelial cell (APC)
Confidence

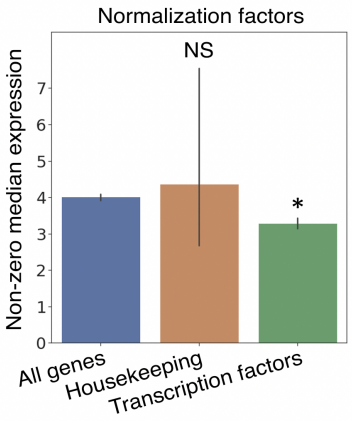

a
Confidence
-4 -2 0 2 4 6 8
Predicted
Actual
Kidney cell types
Endothelial cell (EC) (APC)
Epithelial cell
Fetal epithelial progenitor
Loop of Henle
Ureteric bud cell
Intercalated cell
Fetal mesenchymal progenitor
Intermediated cell
Proximal tubule progenitor
Macrophage
EC
Smooth muscle cell
Fetal stromal cell
Dendritic cell
T cell
b
Confidence
-2 0 2 4 6
Predicted
Actual
Liver cell types
Erythroid progenitor
Erythroid cell
Sinusoidal EC
Macrophage
Dendritic cell
Monocyte
Neutrophil
T cell
B cell (plasmocyte)
Neutrophil
B cell
CB CD34+
c
Predicted
Actual
Blood cell types
Antigen presenting cell
Monocyte
T cell
Neutrophil
Erythroid cell
Erythroid progenitor
Dendritic cell
B cell (plasmocyte)
Neutrophil
B cell
Confidence
8 6 4 2 0 -2 -4
d
Predicted
Actual
Spleen cell types
B cell (plasmocyte)
B cell
T cell
Neutrophil
Macrophage
EC (APC)
Confidence
-4 -2 0 2 4 6 8
e
Predicted
Actual
Brain cell types
Fetal neuron
Fetal mesenchymal progenitor
Macrophage
Erythroid cell
Fetal endocrine cell
Fetal epithelial progenitor
Confidence
-2 0 2 4 6
f
Predicted
Actual
Placenta cell types
Fibroblast
Epithelial cell
Macrophage
Confidence
-4 -2 0 2 4
g
Actual
cells = 6525
embedding dimensions = 256
Lung cell types
Smooth muscle cell
AT2 cell
Fetal mesenchymal progenitor
Macrophage
M2 macrophage
Endothelial cell (EMT)
T cell
Endothelial cell (APC)
Mast cell
B cell (plasmocyte)
Fetal epithelial progenitor
Dendritic cell
Proliferating T cell
Endothelial cell
Monocyte
Basal cell
Embedding dimension activation
0 0.5 1

Furthermore, the benefit of the novel gene ranking strategy is not evaluated.

Authors' response: We thank the reviewer for this discussion. Below we address 1) the normalization strategy that we apply prior to the gene ranking and 2) the rank value encoding of each transcriptome.

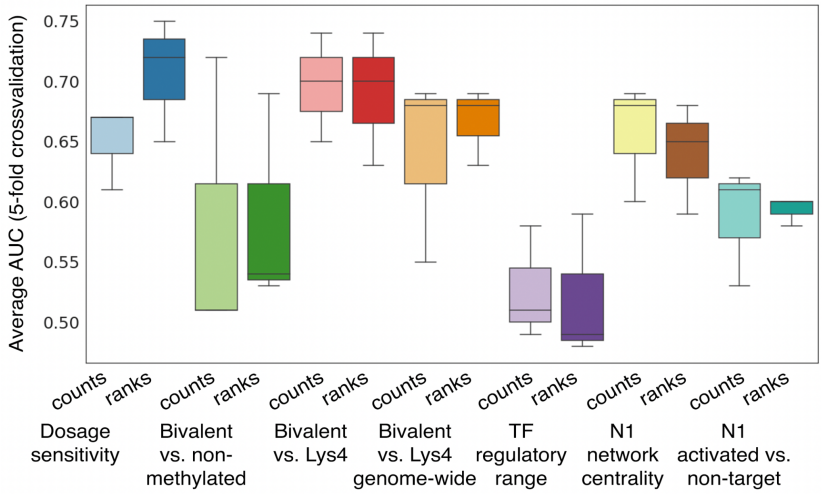
1) Normalization strategy:
We quantified the normalization factors for housekeeping genes and transcription factors compared to all genes to better characterize the effects of the normalization (below and in Extended Data Fig. 1c). Transcription factors are normalized by a statistically significantly lower factor (resulting in higher prioritization in the rank value encoding) compared to all genes. Housekeeping genes on average show a trend of a higher normalization factor (resulting in deprioritization in the rank value encoding) compared to all genes. Thus, normalization serves to prioritize genes like transcription factors that are important to informing the model about network dynamics but would be deprioritized in the rank value encoding if not normalized by their non-zero median value of expression.

Manuscript changes: Addition of Extended Data Fig. 1c (below):



2) Rank value encoding:
Attention-based models are powerful model architectures that have revolutionized the natural language understanding[5,8] field, and there have since been broad efforts to find ways to represent other data types as sequences in order to leverage these powerful model architectures to advance additional fields including image, sound, and video analysis[4,13,14]. We demonstrate in this work that attention-based models can be used with rank value encodings rather than the current standard positional encoding. This knowledge opens the opportunity to leverage these powerful model architectures for data in any field that can be represented as ranks rather than constricting their usage to sequence-based data types. As such, the major benefit of using a rank value encoding is that it allows us to leverage the powerful self-attention model architecture to gain insights in transcriptional networks, which have no inherent sequence-based properties. We have added discussion about this point to the manuscript to further clarify this.

UMI count data cannot be standardly directly used as an input to attention-based models, but when comparing count versus rank data using random forests, support vector machines, and logistic regression in each of the downstream applications we tested, there was no significant advantage to using the count data, suggesting the relevant information is sufficiently represented in the rank value encoding (AUC comparison summarized below, F1 score comparison in Extended Data Fig. 7-9). When applying Geneformer to the rank value encodings, the results outperform these alternative methods that standardly use count data.



(all counts vs. ranks comparisons not significant ($p>0.05$) by Wilcoxon rank sum test)

Manuscript changes: Addition of Extended Data Fig. 7-9 demonstrating confusion matrices and F1 score for alternative methods on ranks vs. counts and discussion as follows:

"The advent of self-attention has revolutionized the NLU field, and there have since been broad efforts to represent other data types as sequences in order to leverage these powerful attention-based model architectures to advance fields including image, sound, and video analysis. In this work, we demonstrate that attention-based models can be applied to rank value encodings rather than the current standard positional encodings. This opens the opportunity to leverage these powerful model architectures for data in any field that can be represented as ranks rather than only sequence-based data types."

C. Data & methodology: validity of approach, quality of data, quality of presentation.

The methodology is explained well and is sounds.
However, a key issue is that the presented approach, which is pre-trained on large datasets and then fine-tuned for a specific task on smaller datasets, is only compared to simple ML methods that are trained solely on the smaller, task-specific datasets (e.g., SVM, linear regression, or the proposed DNN, but without the pre-training on the large dataset). Importantly, the proposed approach is not compared to any other transfer learning methods, while many have been proposed in the literature for the study of single

cell expression data (e.g., see point B). Thus, while the study further sheds light on the benefit of transfer learning to study single cell data, it does not allow for assessing the quality of the proposed DNN architecture / transfer learning strategy.

We thank the reviewer for this question; please see reviewer #2 comment #2 for further discussion.

Furthermore, in their approach, the authors make many methodological choices, e.g., using an attention based DNN with 6 encoder units (each composed of a self-attention layer and of a feed forward layer), using ranks instead of gene expression values, using 512 dimensions for embedding the cells, using precisely 15% of randomly chosen genes in the masked learning objective for the pre-training step, etc… While the intuitions / reasoning behind some of these choices are given, the authors do not present the results showing their benefits.

Authors' response: We thank the reviewer for this question. As a general principle, the goal is to train the deepest attention-based model architecture for which there is sufficient data to train. This approach has been established to yield the greatest predictive potential in other informational fields including NLU, computer vision, and mathematical problem-solving[4]. The width-to-depth aspect ratio was chosen based on ratios found to be effective in analogous attention-based models[5], and this ratio was accordingly scaled to the maximum depth achievable based on our pretraining data quantity and available GPU resources. We also maximized the amount of context considered by the model with full attention, which is considerably larger than standard language models and contributes to the compute required for pretraining given the quadratic memory and time complexity of full attention. We implemented recent advances in distributed GPU training[6,7] to allow efficient pretraining on the large-scale corpus within the available GPU resources.

In the future though, larger pretraining corpuses and further advances in GPU hardware and distribution algorithms will likely allow training of deeper Geneformer models, which may open opportunities to achieve meaningful predictions in even more elusive tasks with increasingly limited task-specific data. Additionally, the input size of Geneformer fully represented 93% of the rank value encodings in Genecorpus-30M, so based on the number of genes standardly detected in each cell using the current scRNA-seq technology, we were able to take advantage of the predictive benefits of full attention (rather than sparse attention) despite the quadratic memory and time complexity. However, future advancements in scRNA-seq technology or reduced sequencing costs may allow detection of larger numbers of genes per cell. At that time, consideration of the greater context size may offer predictive benefits that outweigh the benefits of applying full attention, so future work may design a model to leverage sparse attention to allow consideration of greater context size without the limitations of the quadratic memory and time complexity of full attention.

Furthermore, future advances in GPU hardware and distribution algorithms may allow further optimization of model architectures and training hyperparameters, extensive optimization of

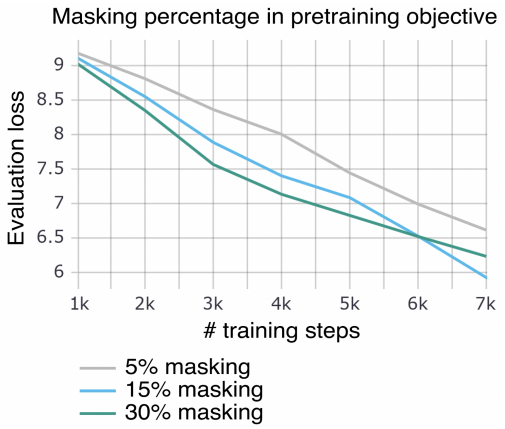
which at the pretraining stage is currently precluded by the large amount of resources needed for pretraining and the fact that optimization with subsampled datasets is known not to generalize to larger datasets due to the scaling laws of these models[4]. Of note, training hyperparameter optimization at the fine-tuning stage is computationally feasible and highly recommended as these hyperparameters will be data- and application-specific. Training hyperparameters were tuned for the cardiomyopathy disease modeling fine-tuning application as described in the Extended Methods ("Disease Modeling" section).

In terms of the rank value encoding, please see reviewer #2 comment #2 for further discussion.

Regarding the masking percentage, 15% was chosen because it is the standard percentage used for masked learning objectives in other informational fields[5]. However, we repeated pretraining on a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for evaluation, and found that 15% masking had marginally lower evaluation loss compared to 5% or 30% masking (below and Extended Data Fig. 1e). We added commentary to the manuscript regarding the above discussion.

Manuscript changes:

- Addition of Extended Data Fig. 1e (below):



- Addition of text regarding parameter choice:

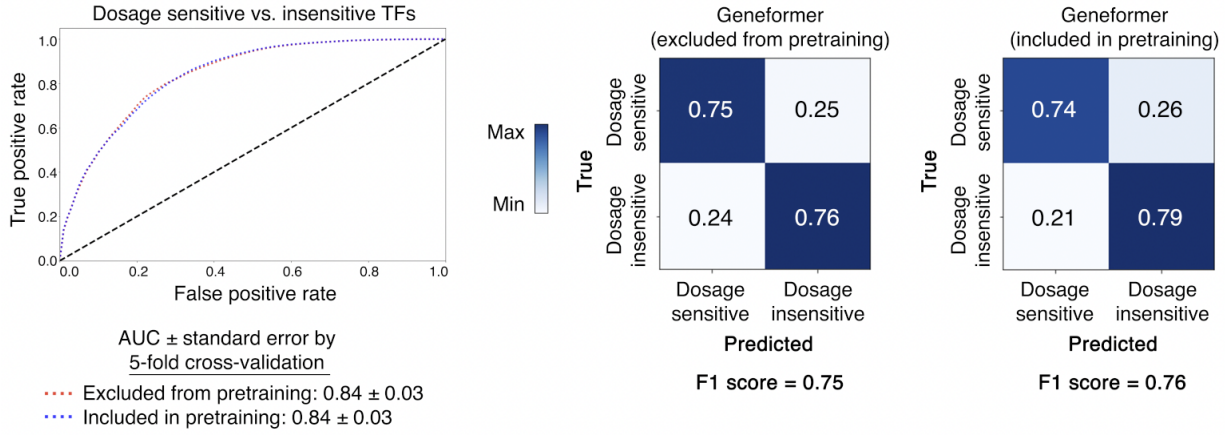
"Depth was chosen based on the maximum depth for which there was sufficient data to pretrain as it has been established that this approach yields the greatest predictive potential in other informational fields including NLU, computer vision, and mathematical problem-solving. Additionally, we maximized the amount of context (input size) considered by the model with full attention based on the number of genes standardly detected in each cell within the pretraining corpus."

As I am not familiar with the usage of the single cell expression data used in the study, I cannot comment on the filtering strategy used by the authors. From a pure ML point of

view, it would be nice if the authors could report / comment on the potential overlap between the large pre-training dataset (from Geneformer-30M) and the smaller, task specific datasets used in the downstream analyses, which could lead to 'data leakage'.

Authors' response: We thank the reviewer for raising this important point. The goal of the large-scale pretraining is to present to the model as many observations as possible to gain the best understanding of gene network dynamics. The fundamental knowledge gained from this large-scale pretraining can then be democratized to a vast array of downstream tasks. It is certainly possible that there could be overlap between the large-scale corpus used for pretraining and smaller task-specific datasets utilized by the user for fine-tuning applications. This is analogous to large-scale pretrained models in the natural language understanding field where the large amount of text used for pretraining could potentially overlap with text utilized by users for fine-tuning applications. However, the pretraining is accomplished using a masked learning objective, which has been shown in other informational fields[5,8] to improve generalizability of the foundational knowledge learned during pretraining for a wide range of downstream fine-tuning objectives. Although beneficial for pretraining, this learning objective is not one that would be practically used for downstream applications. Because the fine-tuning applications are trained on classification objectives that are completely separate from the masked learning objective, whether or not task-specific data was included in the pretraining corpus is not relevant to the classification predictions. Predicting the class of genes is a completely separate learning domain than predicting the identity of the masked genes within a cell. To demonstrate this, we repeated pretraining on a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for the downstream application. We then fine-tuned the pretrained model to distinguish dosage sensitive vs. insensitive transcription factors using either the 10,000 held-out cells or 10,000 cells included in the subsampled pretraining corpus. The performance was equivalent, as shown in Extended Data Fig. 1f.

Manuscript changes: Addition of Extended Data Fig. 1f (below):



D. Appropriate use of statistics and treatment of uncertainties

Statistics are properly used.

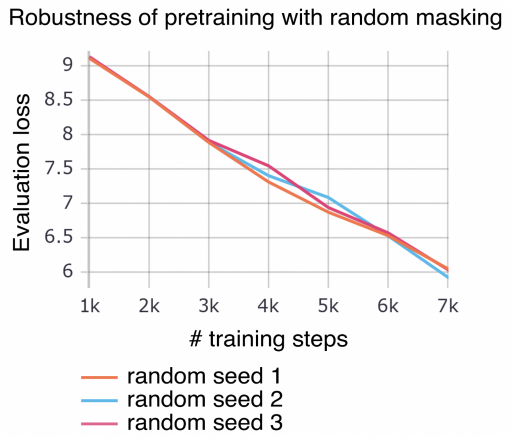
E. Conclusions: robustness, validity, reliability

The robustness of the fine-tuning step is properly evaluated with 5-fold cross validation.

My main concern here is the robustness of the pre-training step, which is not evaluated. In particular, the pre-training process, which is the most computationally intensive step, is heuristically done using a masked learning objective, in which 15% of randomly chosen expressed genes per sample are masked and used as targets for the training process. Due to the randomness of this training process, a robustness analysis is needed.

Authors' response: We thank the reviewer for this suggestion. We repeated the pretraining with a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for evaluation. We conducted 3 trials, each with a different random seed for the 15% masking. The evaluation loss in each trial was essentially equivalent, indicating robustness of the 15% random masking procedure (below and Extended Data Fig. 1d).

Manuscript changes: Addition of Extended Data Fig. 1d (below):



F. Suggested improvements: experiments, data for possible revision

To better position their approach, the authors should include a review of the literature on the transfer learning approaches that have been applied to the analysis of single cell expression datasets.

Based on this review, the authors should compare the performance of their proposed Geneformer to those of relevant transfer learning approaches for single cell expression data analysis from the literature (or at least, they should explain why the comparison is not possible).

We thank the reviewer for this suggestion; please see reviewer #2 comment #2 for further discussion.

The robustness of Geneformer's pre-training step should be evaluated in a cross-fold experiment (evaluating how the ability of Geneformer to boost downstream analysis is impacted by the randomness of the masked learning strategy).

We thank the reviewer for this suggestion; please see reviewer #2 comment E.

The different methodological choices should be evaluated to show their contributions; e.g., the results obtained by using the proposed gene ranking strategy should be evaluated against those that are obtained when using traditional (normalized, batch processed) gene expression values.

We thank the reviewer for this comment. In terms of the rank value encoding, please see reviewer #2 comment #2 for further discussion. In terms of other methodological choices, please see reviewer #2, comment C.

G. References: appropriate credit to previous work?

The authors do not mention / cite any previous work on transfer learning applied on single cell data, while many approaches have been proposed (e.g., the list in point B above).

We thank the reviewer for this suggestion; please see reviewer #2 comment #2 for further discussion.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Given the above-mentioned problems of lack of literature survey on transfer learning applied to single cell data and the lack of comparison with such methods, the paper (including abstract, introduction, and conclusion) does not clearly / objectively present the context of the paper.

We thank the reviewer for this suggestion; please see reviewer #2 comment #2 for further discussion.

Referee Three:

Theodoris et al. developed a computational tool called Geneformer which utilises a compilation of 30M publicly available large-scale single-cell transcriptomes from various tissue and cell types, labelled Genecorpus-30M. Geneformer is pre-trained using a transformer encoder network on Genecorpus-30M to identify the context-specific gene regulatory and expression hierarchies. The authors demonstrate that gene-former is context-aware and is able to characterise network hierarchy through the neural network structure. With fine-tuning using limited data specific to the task, geneformer was capable of predicting known regulators of cell identity, dosage-sensitive genes, bivalency, downstream targets, hierarchy of gene networks and transcription factor cooperativity. The authors further use geneformer to model in-silico deletion to identify disease and cell-identity related genes and screen for potential druggable candidates. Given its broad utility even with limited input data, its applicability in demonstrated and many related subjects would be enormous. However, we have found major flaws and limitations of the study, mainly lack of new biological insight extracted from predictions as well as algorithmic and conceptual advances in the computational method and analysis results are not sufficient to support conclusions.

1. Absence of new biological insight in predictions
All predictions in the study were validated only with previously published literature (i.e. recovery of what is already known). While this is still a valid approach to demonstrate success of Geneformer as a computational tool, it would be much desirable if new predictions were followed up with experimental validation. Importantly, this means that the study did not sufficiently demonstrate discovery capacity of Geneformer. In the revision, it would be great to see if authors can follow up with a few predictions, for example in silico deletion of predicted genes influence downstream gene network. This establishes essential causal relationship between the targeted gene and its biological effect, which lacks in the current study.

Authors' response: We thank the reviewer for this suggestion.

We performed experimental validation to determine whether experimentally targeting one of the top novel dosage-sensitive gene candidates in cardiomyocytes, *TEAD4*, would damage the function of cardiomyocytes. Indeed, we found that CRISPR-mediated knockout of *TEAD4* in iPSC-derived engineered cardiac microtissues caused a significant reduction in their ability to generate contractile stress (defined as contractile force per unit area). TEAD4 is a transcription factor involved in the Hippo signaling pathway, and future work is warranted to further examine its role in cardiac development. This experimental validation supports the utility of Geneformer for driving novel biological insights in human development.

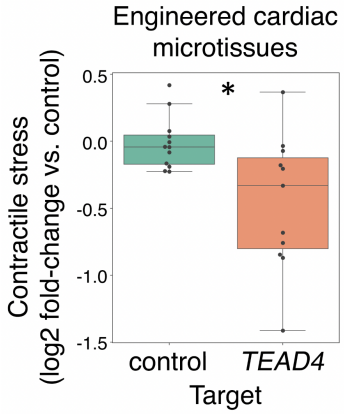
Additionally, we performed experimental validation to determine whether inhibition of candidate therapeutic targets for dilated cardiomyopathy predicted by Geneformer's in silico treatment approach could improve cardiomyocyte function in an experimental model of the disease.

Titin (*TTN*) truncating mutations are the leading cause of dilated cardiomyopathy in humans and are found in ~20% of affected patients[9]. Induced pluripotent stem cell (iPSC)-derived cardiac microtissues harboring a truncating variant ($TTN^{+/-}$) in the A-band are known to exhibit reduced contractile force compared to isogenic $TTN^{+/+}$ controls[10].

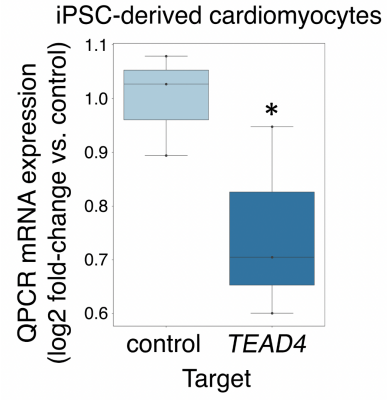
With this in mind, we tested whether inhibition of Geneformer-predicted therapeutic targets would have a functional benefit in the $TTN^{+/-}$ cardiac microtissues. Strikingly and as predicted, CRISPR-mediated knockout of both *GSN* and *PLN* in the $TTN^{+/-}$ cells significantly improved the contractile stress of the $TTN^{+/-}$ engineered cardiac microtissues, validating these genes as promising candidate therapeutic targets for this disease (Fig. 6f, Extended Data Fig. 10e-f). Knockout of *ESRRG* also showed a trend of improving contractile stress, while the fourth target we tested, *HMGB1*, had no effect. These findings provide experimental validation in support of the utility of Geneformer as a tool for discovery of candidate therapeutic targets in human disease.

Manuscript changes:

- Addition of Fig. 2e and Extended Data Fig. 7b (below):

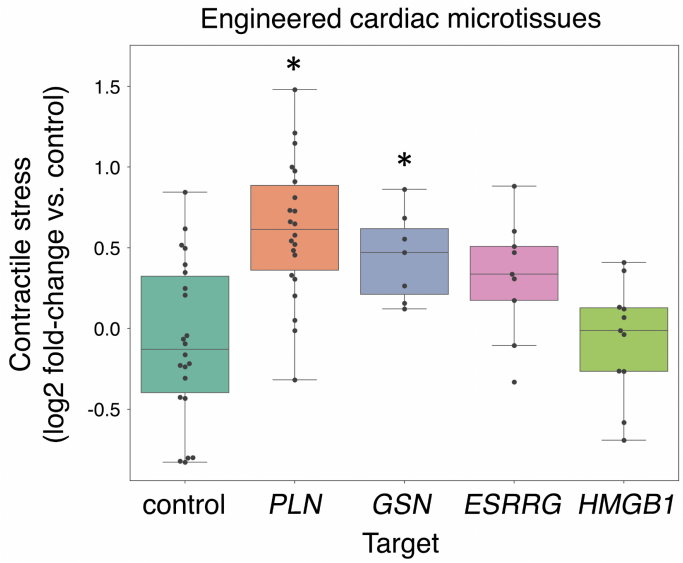


(*WT* + control n=12, *WT* + CRISPR guides targeting knockout of *TEAD4* n=11; *p<0.05 by Wilcoxon rank sum test)

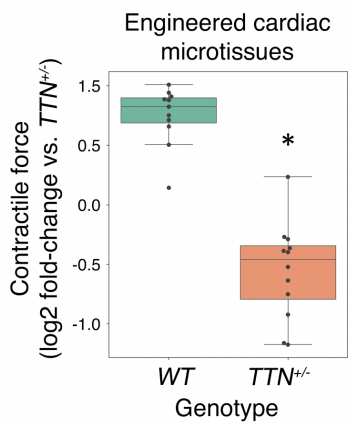


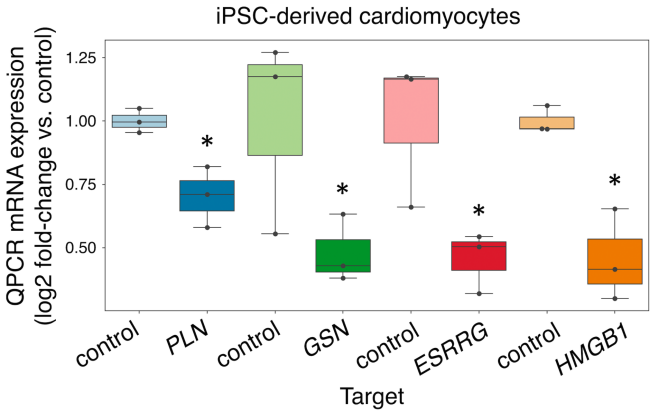
(n=3, *p<0.05 by t-test)

- Addition of Fig. 6f and Extended Data Fig. 10e-f (below):



(*TTN*$^{+/-}$ + control n=22, *TTN*$^{+/-}$ + CRISPR guides targeting knockout of *PLN* n=22, *GSN* n=7, *ESRRG* n=9, and *HMGB1* n=11; *p<0.05 by Wilcoxon rank sum test, BH-corrected)





(*left*: *WT* n=11, *TTN*$^{+/-}$ n=12, *p<0.05 by Wilcoxon rank sum test; *right*: n=3, *p<0.05 by t-test)

- Addition to Extended Data Table 11 of predicted beneficial effect of in silico deletion or activation of genes in cardiomyocytes from dilated cardiomyopathy.

- Addition of Extended Data Table 14: gene set enrichments of dilated cardiomyopathy candidate therapeutic targets from in silico treatment analysis by in silico deletion.

- Addition of results to text:
"We tested whether experimentally targeting one of the top novel dosage-sensitive gene candidates, *TEAD4*, would damage the function of cardiomyocytes. Indeed, we found that

CRISPR-mediated knockout of *TEAD4* in iPSC-derived engineered cardiac microtissues caused a significant reduction in their ability to generate contractile stress (defined as contractile force per unit area) (Fig. 2e, Extended Data Fig. 7b). TEAD4 is a transcription factor involved in the Hippo signaling pathway[28], and future work is warranted to further examine its role in cardiac development."

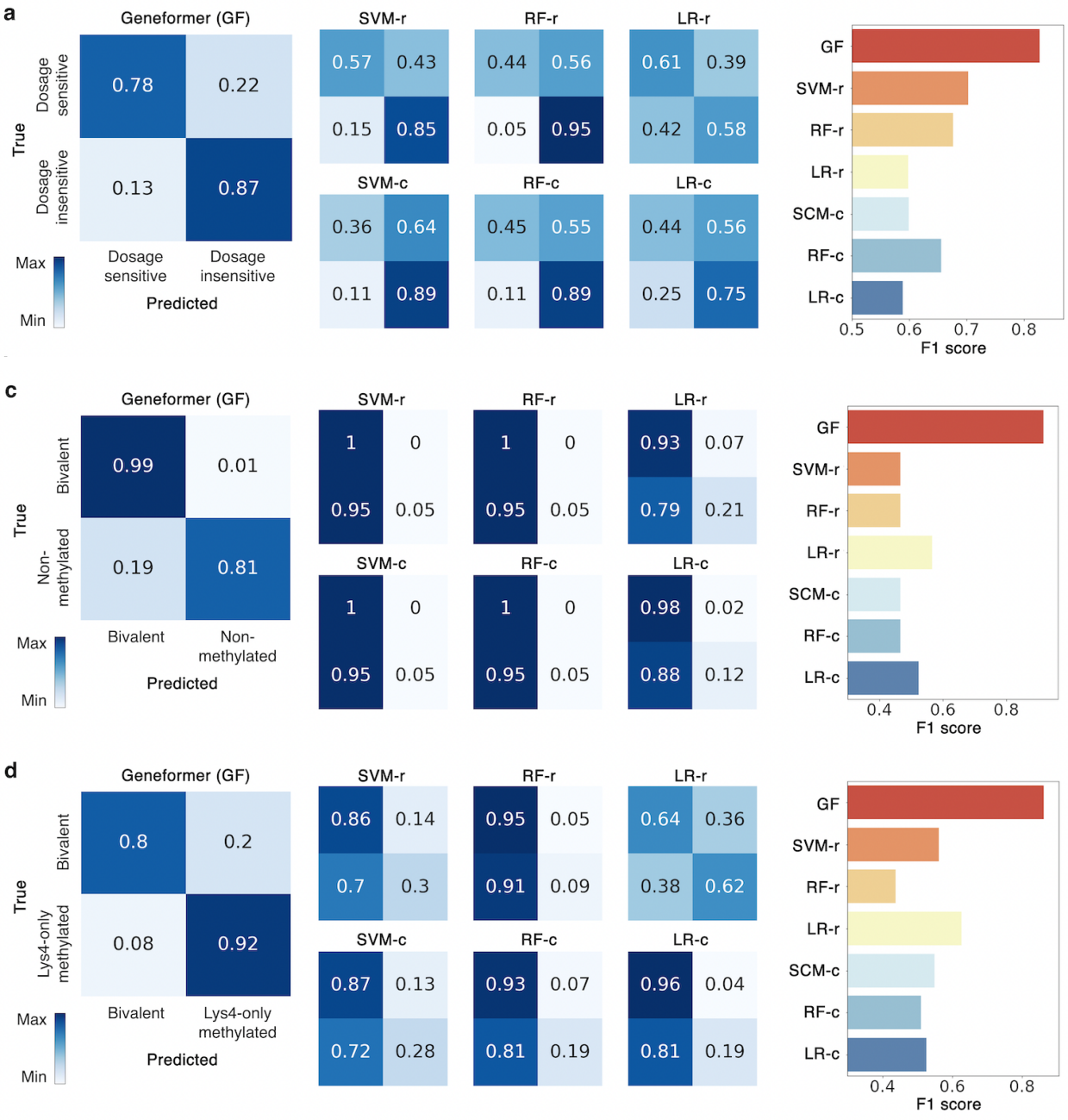
"Finally, we performed in silico treatment analysis in cardiomyocytes from dilated cardiomyopathy patients to determine whether inhibition or activation of specific pathways would shift the cell embeddings back towards the non-failing heart state (Extended Data Table 11, 14). We then performed experimental validation to determine whether inhibition of candidate therapeutic candidates for dilated cardiomyopathy predicted by our in silico treatment approach could improve cardiomyocyte function in an experimental model of the disease. *Titin* (*TTN*) truncating mutations are the leading cause of dilated cardiomyopathy in humans and are found in ~20% of affected patients[49]. iPSC-derived cardiac microtissues harboring a truncating variant ($TTN^{+/-}$) in the A-band are known to exhibit reduced contractile stress compared to isogenic $TTN^{+/+}$ controls[50] (Extended Data Fig. 10e). Therefore, we tested whether inhibition of Geneformer-predicted therapeutic targets would have a functional benefit in the $TTN^{+/-}$ cardiac microtissues.

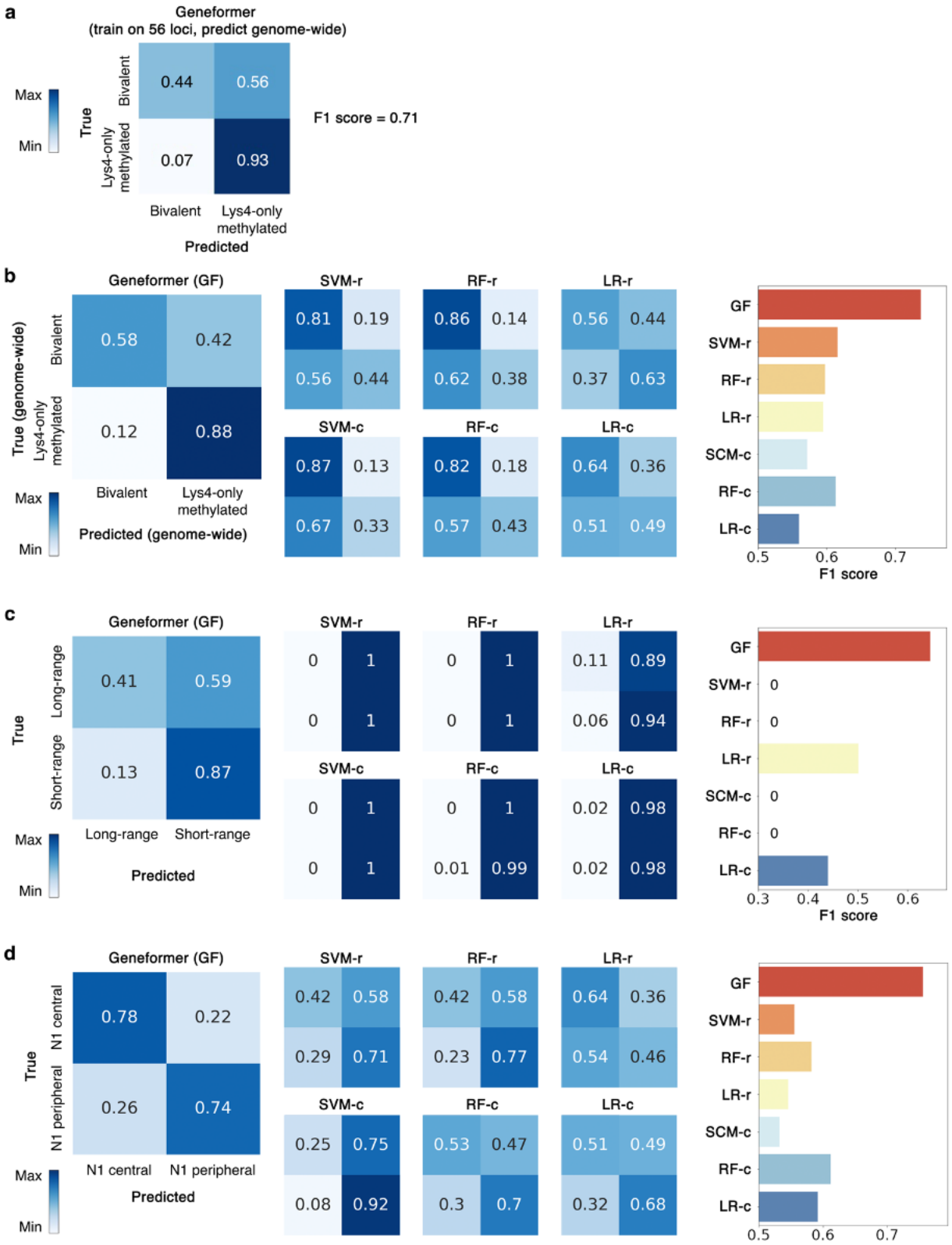
Strikingly and as predicted, CRISPR-mediated knockout of both *GSN* and *PLN* in the $TTN^{+/-}$ cells significantly improved the contractile stress of the $TTN^{+/-}$ engineered cardiac microtissues, validating these genes as promising candidate therapeutic targets for this disease (Fig. 6f, Extended Data Fig. 10e-f). Knockout of *ESRRG* also showed a trend of improving contractile stress, while the fourth target we tested, *HMGB1*, had no effect. These findings provide experimental validation in support of the utility of Geneformer as a tool for discovery of candidate therapeutic targets in human disease."

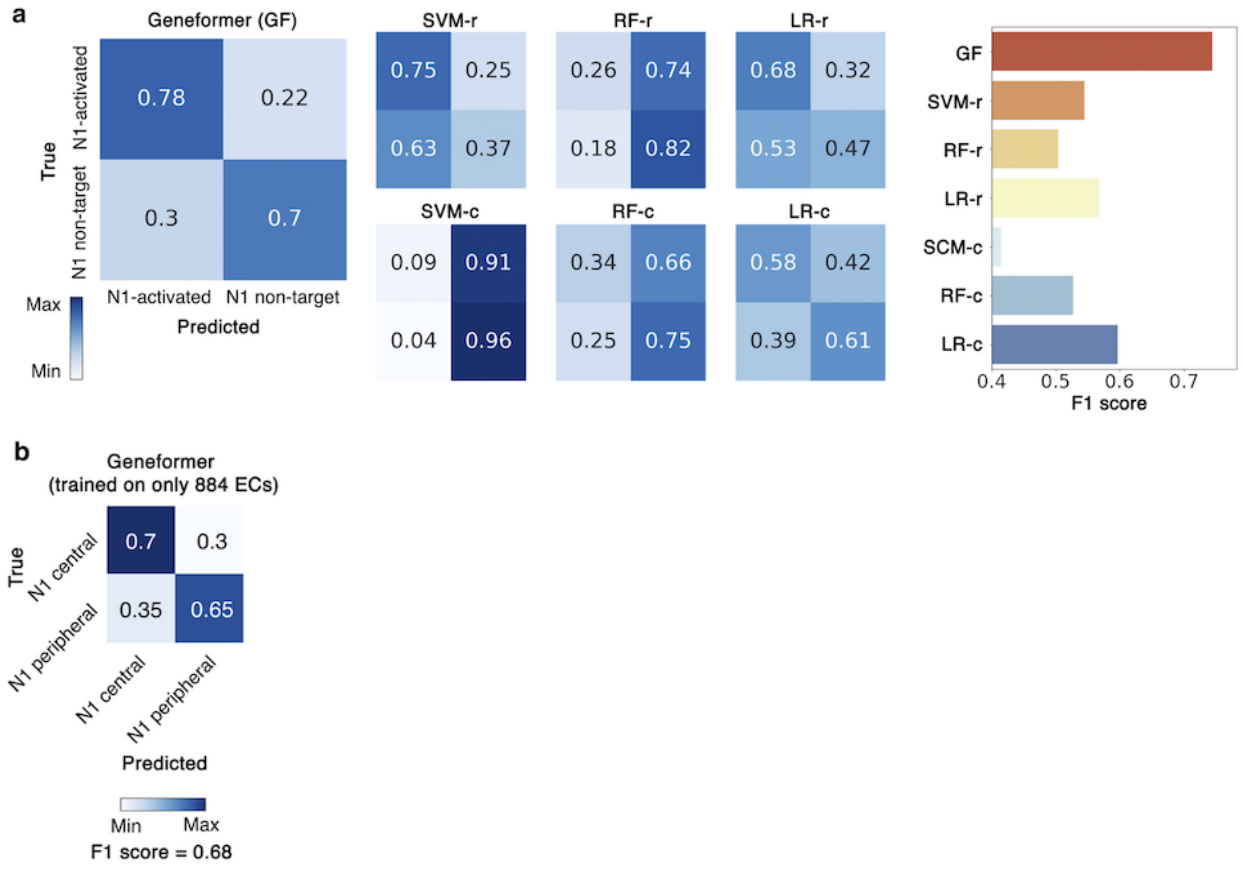
2. Suboptimal comparisons for computational algorithm and output
A few limitations related to performance testing and output interpretation. First, performance analysis of Geneformer largely focused on recovery of true positives vs. false positives in ROC. ROC curves can be misleading if data are imbalanced but it is not clear how balanced (or imbalanced) the testing data. Precision-recall curve (PRC) helps address this issue especially huge imbalance is present in the data – which is possible given positives are much rarer than negatives in these contexts. This may not be necessary but at least would be ideal to declare if data imbalance exists in test sets.

Authors' response: We thank the reviewer for this important point. We added both confusion matrices and F1 score for each of the analyses previously presented as ROC curves to more clearly demonstrate the performance for both classes (Extended Data Fig. 7-9). Geneformer's F1 score was higher than the alternative methods in all cases.

Manuscript changes: Addition of Extended Data Fig. 7a, 7c-d, 8, and 9a-b (below):


a
Geneformer (GF)
SVM-r
RF-r
LR-r
SVM-c
RF-c
LR-c
True
Dosage sensitive
Dosage insensitive
Predicted
Max
Min
0.78
0.22
0.13
0.87
0.57
0.43
0.15
0.85
0.44
0.56
0.05
0.95
0.61
0.39
0.42
0.58
0.36
0.64
0.11
0.89
0.45
0.55
0.11
0.89
0.44
0.56
0.25
0.75
GF
SVM-r
RF-r
LR-r
SCM-c
RF-c
LR-c
0.5
0.6
0.7
0.8
F1 score
c
Bivalent
Non-methylated
0.99
0.01
0.19
0.81
1
0
0.95
0.05
1
0
0.95
0.05
0.93
0.07
0.79
0.21
1
0
0.95
0.05
1
0
0.95
0.05
0.98
0.02
0.88
0.12
0.4
0.6
0.8
d
Lys4-only methylated
0.8
0.2
0.08
0.92
0.86
0.14
0.7
0.3
0.95
0.05
0.91
0.09
0.64
0.36
0.38
0.62
0.87
0.13
0.72
0.28
0.93
0.07
0.81
0.19
0.96
0.04
0.81
0.19

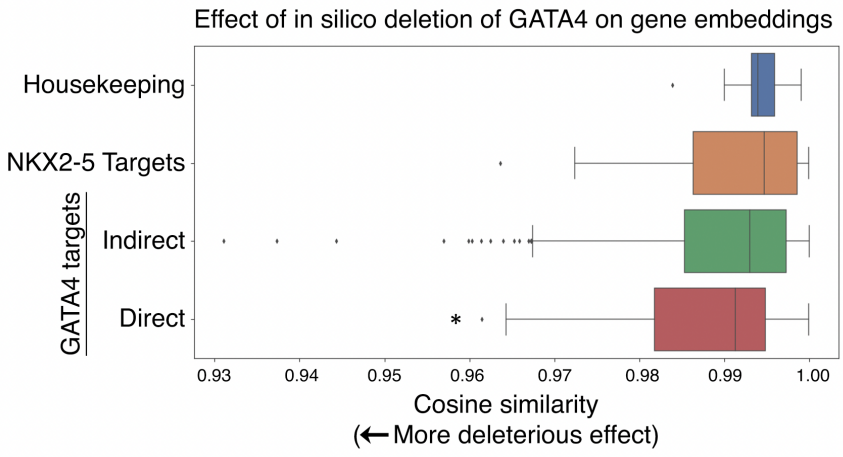

a
Geneformer
(train on 56 loci, predict genome-wide)
Max
Min
True
Bivalent
Lys4-only methylated
0.44
0.56
0.07
0.93
Bivalent
Lys4-only methylated
Predicted
F1 score = 0.71
b
Geneformer (GF)
True (genome-wide)
0.58
0.42
0.12
0.88
Predicted (genome-wide)
SVM-r
0.81
0.19
0.56
0.44
RF-r
0.86
0.14
0.62
0.38
LR-r
0.56
0.44
0.37
0.63
SVM-c
0.87
0.13
0.67
0.33
RF-c
0.82
0.18
0.57
0.43
LR-c
0.64
0.36
0.51
0.49
GF
SVM-r
RF-r
LR-r
SCM-c
RF-c
LR-c
0.5
0.6
0.7
F1 score
c
Geneformer (GF)
True
Long-range
Short-range
0.41
0.59
0.13
0.87
Predicted
SVM-r
0
1
0
1
RF-r
0
1
0
1
LR-r
0.11
0.89
0.06
0.94
SVM-c
0
1
0
1
RF-c
0
1
0.01
0.99
LR-c
0.02
0.98
0.02
0.98
GF
SVM-r
0
RF-r
0
LR-r
SCM-c
0
RF-c
0
LR-c
0.3
0.4
0.5
0.6
F1 score
d
Geneformer (GF)
True
N1 central
N1 peripheral
0.78
0.22
0.26
0.74
Predicted
SVM-r
0.42
0.58
0.29
0.71
RF-r
0.42
0.58
0.23
0.77
LR-r
0.64
0.36
0.54
0.46
SVM-c
0.25
0.75
0.08
0.92
RF-c
0.53
0.47
0.3
0.7
LR-c
0.51
0.49
0.32
0.68
GF
SVM-r
RF-r
LR-r
SCM-c
RF-c
LR-c
0.5
0.6
0.7



Second, effect analysis of in silico deletion of GATA4 & TBX5 (Fig 5b) was compared between known co-bound targets vs housekeeping genes. This analysis is incomplete and potentially misleading as the comparison did not show its effect on other TF targets. Without knowing this, we cannot conclude that Geneformer is capable of predicting correct network hierarchy specific to a given context, hence context-aware. Is the effect broader over TFs or really specific for these targets? This discrepancy needs to be clearly answered by for example comparing against randomly chosen TF targets.

Authors' response: We thank the reviewer for suggesting this control to better evaluate the specificity of the predicted targets. In Fig. 5a, in silico deletion of GATA4 was statistically significantly deleterious to direct target genes, rather than indirect targets or housekeeping genes. To confirm that the effect was specific to GATA4 direct targets and not other direct transcription factor targets as suggested by the reviewer, we also tested direct targets of NKX2-5, another important transcription factor regulating cardiomyocyte function and development. We found that NKX2-5 direct targets were not statistically significantly impacted by GATA4 in silico deletion compared to housekeeping genes and were less impacted than GATA4 indirect targets (Fig. 5a and below).

Manuscript changes: Addition of NKX2-5 target control group to Fig. 5a (below):



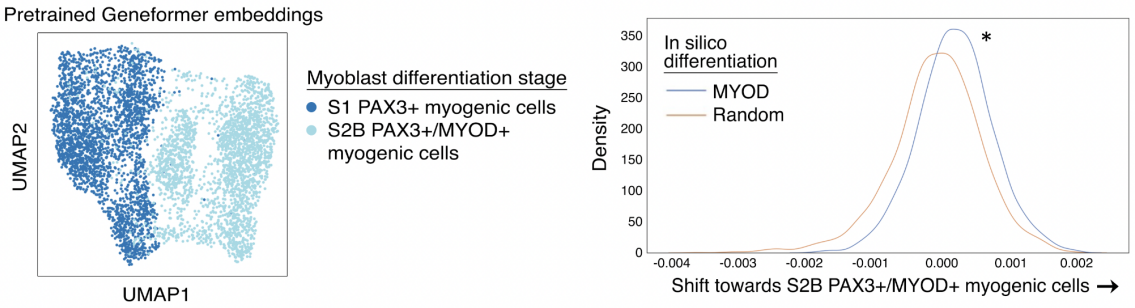
3. Magnitude of in-silico deletion effects
Throughout the paper, in-silico deletion of genes is used to demonstrate context-awareness and model disease and/or potential drugs. Firstly, Effect analysis of in silico expression of OKSM factors (Extended Data Fig 1C) which was used to demonstrate the context-awareness of Geneformer was noted to significantly shift gene embeddings toward an iPSC state. However, a comparison point of gene-embeddings in an iPSC state should be provided to show the magnitude of the shift. Furthermore, the gene embeddings after artificially expressing OKSM could be compared to the gene embeddings from the iPSC reprogramming dataset presented in Extended Data Fig 1B. Demonstration of context-awareness could also be performed with additional reprogramming factors, for instance MyoD.

Secondly, in the in-silico therapy of cardiomyopathy, gene candidates were identified which shifted cells from a non-failing state to a cardiomyopathy state, and then druggable candidates which shifted the cells to a non-failing state. The magnitude of these shifts is unclear in the results – it would be beneficial to show how much a gene's deletion shifts cell embeddings toward the cardiomyopathy or non-failing state respectively.

Authors' response: We thank the reviewer for this point of clarification. The in silico perturbation compares the embeddings of genes in the starting cell with a cell context that is otherwise identical except for the perturbed gene's rank. The end state cells define the coordinates of the end point to specify the goal direction of the embedding vector shift. However, the embeddings of genes in the end state will always be more different than the embeddings of genes in the perturbed cell because the context in the end cell is the biological end state of the perturbation whereas the perturbed cell is encoded as only the first step in the downstream changes that would biologically occur in response to the perturbation. For this reason, the magnitude of the shift in response to a gene's in silico perturbation is most appropriately compared to the distribution of shift that occurs in response to in silico perturbation of random genes (as shown in Extended Data Fig. 2c,e).

We thank the reviewer for the suggestion of additional in silico overexpression studies. Guo et al. *eLife* 2022 reported generation of iPSC-derived myoblasts that pass through an initial PAX3+/MYOD- stage (S1) towards the PAX3+/MYOD+ myogenic state (S2B)[11]. We tested in silico differentiation via artificially adding MYOD to the front of the rank value encodings of PAX3+/MYOD- (S1) cells and found the embeddings of the remaining genes significantly shifted towards the MYOD+ (S2B) myogenic cell state (below and Extended Data Fig. 2d-e).

Manuscript changes: Addition of Extended Data Fig. 2d-e and relevant text:



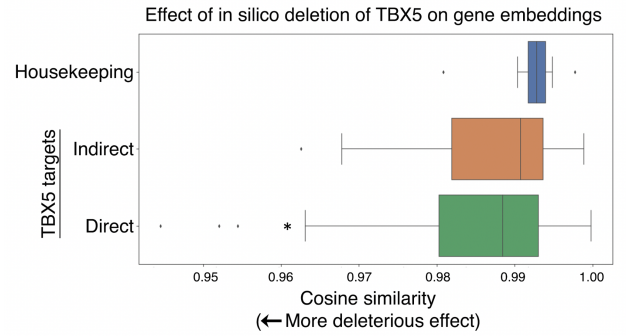
"Embeddings of genes in iPSC-derived myogenic cells showed similar context awareness with in silico differentiation via MYOD (Extended Data Fig. 2d-e)."

Finally, if authors could calculate how faithfully in-silico deletions model in-vitro knock outs on gene embeddings. This could be performed using existing RNA datasets of gene knockouts.

Authors' response: We thank the reviewer for this suggestion. We presented data indicating that in silico deletion of GATA4 appropriately modeled the effects of reduced dosage of GATA4 noted in a previously reported iPSC disease model by RNA-seq. Specifically, in silico deletion of GATA4 had a significantly higher effect on the embeddings of genes known to be most significantly dysregulated by GATA4 variants in a previously reported iPSC disease model of GATA4-related heart defects[1] (Extended Data Fig. 9d). Notably, direct GATA4 targets (as defined by GATA4 binding by ChIP-seq in the iPSC disease model[1]) were significantly more impacted by in silico deletion of GATA4 in fetal cardiomyocytes compared to indirect targets (Fig. 5a).

To further investigate this, we tested a second previously reported iPSC disease model, this time of TBX5 deletions[12]. We tested the effect of in silico deletion of TBX5 on TBX5 targets known to be significantly dysregulated in response to TBX5 deletions in the iPSC disease model. Again, we found that direct TBX5 targets (as defined by TBX5 binding by ChIP-seq in the iPSC disease model[1]) were significantly more impacted by in silico deletion of TBX5 in fetal cardiomyocytes compared to indirect targets (below and Extended Data Fig. 9e).

Manuscript changes: Addition of Extended Data Fig. 9e and relevant text:



"Analogously, in silico deletion of TBX5, another known congenital heart disease gene, in fetal cardiomyocytes more significantly impacted its direct targets (as defined by ChIP-seq) compared to indirect targets and housekeeping genes (Extended Data Fig. 9e)."

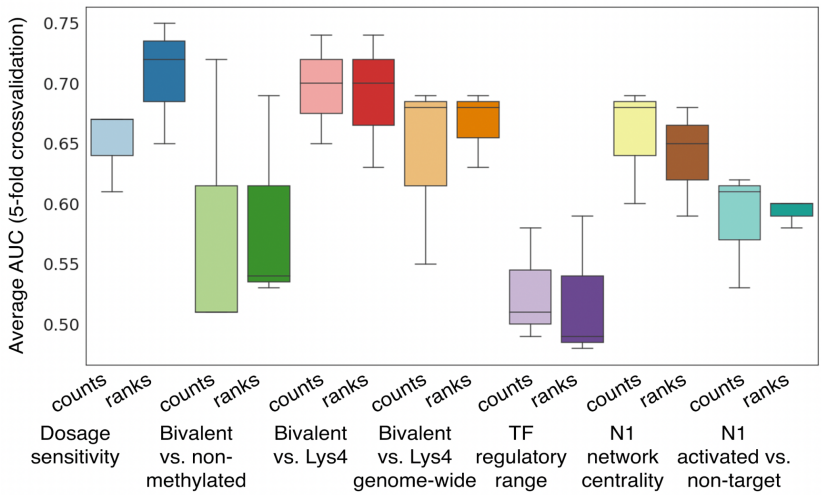
4. The benefits from transforming UMI count data to rank-based matrix need to be justified and technical issues related to using the rank metric with regard to genes sharing the same rank need to be assessed. Here, scRNAseq UMI count data is transformed into rank-based data for inputting into the neural network and the authors claimed that this transformation is a novel contribution of their method, however the benefits of this rather simple preprocessing step need more evidence.

The rank transformation may cause issues because many genes may share the same rank because in a scRNAseq UMI count data, the value range is much less diverse compared to bulk sequencing. The issue of shared ranks is even more common for many genes with 0 UMI count in any given cell. This is a classical issue due to the inherent sparsity of the single cell data. The authors should assess the effect of discretising the UMI data to rank data, with regard to how the data distribution changes and how these changes affect downstream analyses.

Authors' response: We thank the reviewer for this discussion. Attention-based models are powerful model architectures that have revolutionized the natural language understanding[5,8] field, and there have since been broad efforts to find ways to represent other data types as sequences in order to leverage these powerful model architectures to advance additional fields including image, sound, and video analysis[4,13,14]. We demonstrate in this work that attention-based models can be used with rank value encodings rather than the current standard positional encoding. This knowledge opens the opportunity to leverage these powerful model architectures for data in any field that can be represented as ranks rather than constricting their usage to sequence-based data types. As such, the major benefit of using a rank value encoding is that it allows us to leverage the powerful self-attention model architecture to gain insights in transcriptional networks, which have no inherent sequence-based properties. We have added discussion about this point to the manuscript to further clarify this.

UMI count data cannot be standardly directly used as an input to attention-based models, but when comparing count versus rank data using random forests, support vector machines, and

logistic regression in each of the downstream applications we tested, there was no significant advantage to using the count data, suggesting the relevant information is sufficiently represented in the rank value encoding (AUC comparison summarized below, F1 score comparison in Extended Data Fig. 7-8, 9a-b). When applying Geneformer to the rank value encodings, the results outperform these alternative methods that standardly use count data.



(all counts vs. ranks comparisons not significant ($p>0.05$) by Wilcoxon rank sum test)

Regarding the question about genes sharing the same rank, because the rank transformation includes normalization by the non-zero median value of expression of that gene across the entire Genecorpus before rank ordering, it is very unlikely that detected genes would share the same rank. We tested this in an example dataset of 75,451 cells undergoing iPSC to cardiomyocyte differentiation assayed in parallel on the Drop-seq (single-cell) or DroNc-seq (single-nucleus) platform[15] and found no occurrences of any detected genes in any cell sharing the same rank. The undetected genes are not included in the rank value encodings so they are appropriately encoded as the same rank through their absence. This is analogous to how attention-based models are used in natural language understanding, where words that are not found in a sentence are also not included in the positional encoding of that sentence, and the model interprets the sentence in the context of their absence.

Manuscript changes: Addition of Extended Data Fig. 7-9 demonstrating confusion matrices and F1 score for alternative methods on ranks vs. counts and discussion as follows:

“The advent of self-attention has revolutionized the NLU field, and there have since been broad efforts to represent other data types as sequences in order to leverage these powerful attention-based model architectures to advance fields including image, sound, and video analysis. In this work, we demonstrate that attention-based models can be applied to rank value encodings rather than the current standard positional encodings. This opens the opportunity to leverage these powerful model architectures for data in any field that can be represented as ranks rather than only sequence-based data types.”

5. More normalisation is required to provide evidence on how the self attention transformer results in the batch effect removal

In their workflow, for each gene, they use global median expression value for only non-zero value, discarding the cases where the gene has 0 value. This approach does not address the uneven distribution of zero values across genes, cells and contexts. The number of 0 values is also dependent on sequencing depth, with more 0s expected if the depth does not reach a saturation level. Normalisation sequencing depth does not solve this issue. In the extended methods, the authors described normalisation across samples based on sequencing depth (to total transcript count per cell) followed by normalising gene counts using non-zero median expression values across all samples (although not clearly described how this is done) before ranking genes for each cell. Also the aggregation of data sub(sets) among the 30M single cells does not account for cell type information, nor sample-to-sample normalisation is used.

Consequently, the results showing the normalisation effect are not convincing. Extended Figure 1a shows results for assessing batch effect, but the density plots do not show how the embedding comparisons reflect the similarity/dissimilarity in the original values. The authors show that embeddings of gene A and gene B in the same cell are more different than those of gene A vs gene A in the same cell type regardless of the batch effects (cosine similarity). The lower similarity between two genes in the same cell relative to the same gene between two different cells is expected and is not sufficient to suggest that the embedding is context dependent. The authors described that they compared UMAP of cells based on the embeddings with UMAP of original gene expression space. However, the resulting UMAPs are different with a clear reduction in the ability to separate cells in different conditions (extended Figure 2a) . Extended Figure 2f shows some improvement in data integration,
but the UMAPs still suggest suboptimal integration, because cardiomyocyte 1 and 2 are still separated by platforms

Authors' response: We thank the reviewer for raising this discussion. The important comparison in the density plots is not the embeddings of different genes in the same cell compared to the same gene in two different cells; as the reviewer suggests, the difference in cosine similarity here is expected, and the leftmost density curve indicating lower similarity between different genes in the same cell was only included as a positive control. The relevant comparison to the question of batch effects is between the rightmost two density curves, which demonstrate that a given gene's embedding is not significantly affected by batch effect. If there was a significant batch effect on the gene embeddings, we would expect that there would be a lower cosine similarity between a given gene's embedding in cells of different batches compared to the embeddings of that gene in cells of the same batch. However, we do not observe this; the gene embeddings are not statistically significantly different when compared in cells from different batches compared to cells from the same batch. We demonstrate this for three major batch effect categories of sequencing platform, preservation method, and individual patient variability.

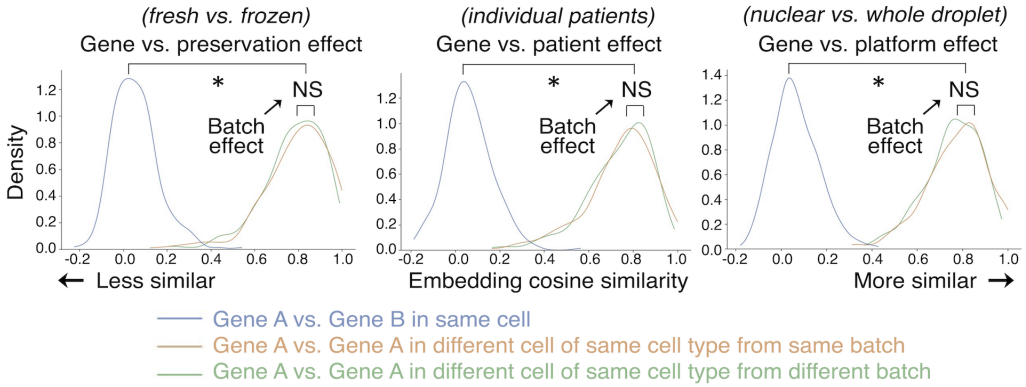
Regarding the cell-level embeddings, as the reviewer noted, these will be necessarily impacted by variable gene detection based on sequencing depth or platform detection rate because the cell embeddings are calculated by integrating the embeddings of the genes detected in that cell. We analyzed data from the iPSC to cardiomyocyte differentiation study to investigate this issue, as the two platforms vary in the number of genes they detect (Extended Data Fig. 4b). The original data was highly impacted by batch, and Geneformer embeddings integrated the data better than the current standard methodologies in the field (Extended Data Fig. 4f). The cardiomyocyte 1 and 2 are necessarily separated by batch because they are unevenly detected in each batch (Extended Data Fig. 4e). Given that the Geneformer model was fine-tuned to distinguish these two cardiomyocyte cell types, they necessarily and appropriately cluster separately in the UMAP, which because of their uneven distribution between batches also necessarily shows a separation of the platforms in these cell subsets.

Regarding the aortic cell embeddings, there is a clear improvement in the integration of individual patients when compared to the original data (Extended Data Fig. 4a). The original data clusters primarily by patient, whereas the Geneformer embeddings cluster primarily by cell type. Of note, these aortic cell UMAPs are based on pretrained Geneformer embeddings; fine-tuning Geneformer to distinguish cells by a particular characteristic (e.g. phenotype) would lead to further separation by that characteristic if desired for that specific downstream application, as illustrated in Extended Data Fig. 4c and Extended Data Fig. 10b. In the cardiomyopathy disease modeling application, the pretrained Geneformer embeddings again improved integration of individual patients compared to the original data. Then, fine-tuning to distinguish cardiomyocytes from the non-failing, hypertrophic cardiomyopathy, and dilated cardiomyopathy states promoted increased separation of these disease states in the cell embedding space.

Overall, the gene embeddings were robust to batch effects caused by sequencing platform, preservation method, and individual patient variability, and the cell embeddings showed improved integration across batches compared to the original data. However, although this is a beneficial outcome, it should be clearly noted that batch effect removal is not a primary goal of the Geneformer model. Incorporation of as much data as possible during pretraining, including observations of datasets from many different batches, significantly boosts predictive potential in downstream applications, as demonstrated for transcription factor dosage sensitivity in Fig. 2b. Performing the pretraining with an entirely self-supervised approach is critical to inclusion of large amounts of data rather than restricting data to that which has accompanying labels such as cell type annotation mentioned by the reviewer. However, for individual downstream applications, users may elect to preprocess their limited task-specific data with batch effect removal or other normalization methods if they find their results are impacted by experimental artifacts in some way. We did not need to perform batch normalization of the cardiomyopathy dataset used for fine-tuning Geneformer for the disease modeling application, but we added a statement in the manuscript to indicate that users may elect to do so at their discretion to specifically address this reviewer's comment.

Manuscript changes:

- Additional labeling in Extended Data Fig. 2a to further clarify the relevant comparison in the density plots to examine the effect of technical batch-dependent artifacts (now indicated by "Batch effect" with arrow as below):



- Addition of commentary on batch integration for fine-tuning applications:

"Of note, although we found that Geneformer was robust to the common batch-dependent technical artifacts that we tested, Geneformer was not designed specifically for scRNA-seq batch integration. As such, users may elect to preprocess their limited task-specific data with alternative batch integration methods prior to using that data for fine-tuning Geneformer towards their downstream task if they find their dataset to be persistently affected by such batch-dependent artifacts."

6. Interpretability of the network needs to be improved. The authors have not adequately described how the attention layers learn the gene and cell context, but rather assume the attention heads in the pretraining step can learn the gene and cell contexts across the 30M transcriptomes. The authors described the analysis of attention weights extracted from attention heads could capture gene regulation network hierarchy and show examples in Figure 4. The concept on how the attention heads capture the gene hierarchy with regards to which gene pays more attention to other genes is described, but figures to prove the model captures this hierarchy is not convincing. For example, the attention for NOTCH1 is higher in layers o, 1 but lower in subsequent layers compared to TCF4 and SOX7 (Figure 4e). This variation from layers to layers and heads to heads appear to be independent of biological context. Such random variation is also observed in Figure 4g.

Authors' response: We thank the reviewer for this discussion. Multiple attention heads are employed in order for the model to learn in an unsupervised manner to pay attention to distinct classes of genes to jointly improve predictions without prior knowledge of any gene's biological function. It is expected that the various layers and attention heads pay attention to different

aspects of the input space to optimize predictions in the given learning objective. To determine whether the pretrained model encoded features of network hierarchy, we analyzed attention weights on classes of genes expected to be biologically influential and found that indeed, specific attention heads significantly attended 1) transcription factors more than other genes, 2) central genes more than peripheral genes within a defined network, and 3) highest ranked genes more than lowest ranked genes in different cell type contexts (Fig. 4e-g, Extended Data Fig. 9c). This indicated that the attention is cell context-aware and consistently encoded network hierarchy in multiple different cell types.

Furthermore, we developed an in silico perturbation approach to analyze gene and cell embeddings, which are the joint interpretable output of the attention weights of the preceding layers. We demonstrated that the pretrained model without any fine-tuning encoded context-specific dosage sensitivity of genes (Fig. 2d), gene network connections (Fig. 5a, Extended Data Fig. 9d-e), and transcription factor cooperativity (Fig. 5b). Furthermore, cell-level embeddings by the pretrained model without any fine-tuning encoded cell context of cells undergoing iPSC reprogramming and myoblast differentiation, and in silico activation of genes known to promote these cell transitions indeed significantly shifted embeddings along the expected trajectory towards iPSCs or myoblasts, respectively (Extended Data Fig. 2b-e).

Mechanistic interpretability of transformer models is an active field of research investigation and interpretability research is generally pursued on much smaller models[16,17] (e.g. 1-2 attention layers, excluding the feed forward neural network component). However, recent work by the Anthropic research group[16,17] noted that there are variable patterns of attention employed by different attention heads and certain types of attention heads only emerge in models with at least two layers. In our analysis of attention heads in each layer, indeed we found consistency in the pattern of attention paid by each head in different cell type contexts.

For example, as discussed above, network centrality-driven attention heads consistently attended to a significantly greater degree the highest ranked genes in each cell's unique rank value encoding in aortic ECs, smooth muscle cells, T cells, and macrophage/monocyte/dendritic cells (which each have different sets of highest ranked genes based on cell type context) (Fig. 4g, Extended Data Fig. 9c). Other attention heads paid attention to a breadth of gene ranks or focused on lowly ranked genes, again consistently in each of the aforementioned cell types analyzed (Fig. 4g, Extended Data Fig. 9c). Overall, there appeared to be a consistent pattern moving through each layer in terms of the heads' attention patterns. Specifically, attention heads in the earliest layers were consistently the most diverse in terms of gene ranks they attended, suggesting the model initially orients to the observed cell state through a joint survey of distinct portions of the input space. The middle layers were most broad in terms of gene ranks they attended, and the final layers were dominated by centrality-driven attention heads that focused on the highest ranked genes that uniquely define each cell state. Although outside the scope of this initial manuscript, Geneformer's application of the transformer architecture to network biology opens an opportunity to further investigate the mechanisms underlying transformer circuits, especially with rank-based inputs.

Manuscript changes:

- Addition of Extended Data Fig. 2d-e and Extended Data Fig. 9e to further analyze the pretrained model's encoding of gene and cell context.

- To address the reviewer's specific comment regarding the prior Fig. 4e and Extended Data Fig. 3a, which plotted attention weights for individual genes within the N1 subcircuit, these plots were meant to serve as a specific example to directly visualize the attention weights of individual genes. We noted that SMAD1, which is more peripheral, was overall less attended than more central N1, SOX7, and TCF4. However, more insight is drawn from rigorously quantifying the attention weights broadly on particular classes of genes as discussed above, so we removed the analysis of the individual genes to improve clarity.

In addition, the interpretation of gene embedding is also questionable. In extended Figure 1 d, why are there a large subset of the 256 latent dimensions that have low embedding values across all genes all cell types?

We thank the reviewer for bringing up this point for clarification. The top plot in Extended Data Fig. 3 (previously Extended Data Fig. 1d) is a plot of the variance of the embedding values, not the embedding values themselves. We analyzed the variance of the embedding values to test whether there are genes with high variance of their embedding values across different contexts (in this case, different aortic cell types) which are therefore context-specific genes, compared to genes with low variance across different contexts that are therefore less context-dependent. NOTCH pathway members, which are known to be highly context-specific, showed higher variance across cell types compared to the housekeeping gene GAPDH. We added a plot of the embedding values themselves, which show variable levels of activation across the different dimensions (below and bottom plot in Extended Data Fig. 3). The activation of specific dimensions associated with each gene, and certain dimensions were commonly activated among gene groups (for example, the dimensions clustering on the left were jointly activated in the NOTCH receptors 1-3).

Another plot of the embedding values themselves that the reviewer may refer to is in Fig. 6c. In this plot, the embedding dimensions encode the identity of the disease states the model was fine-tuned to distinguish and again show variable levels of activation across the different classes. For example, the embedding dimensions that cluster on the right are primarily highly activated in cardiomyocytes from non-failing hearts.

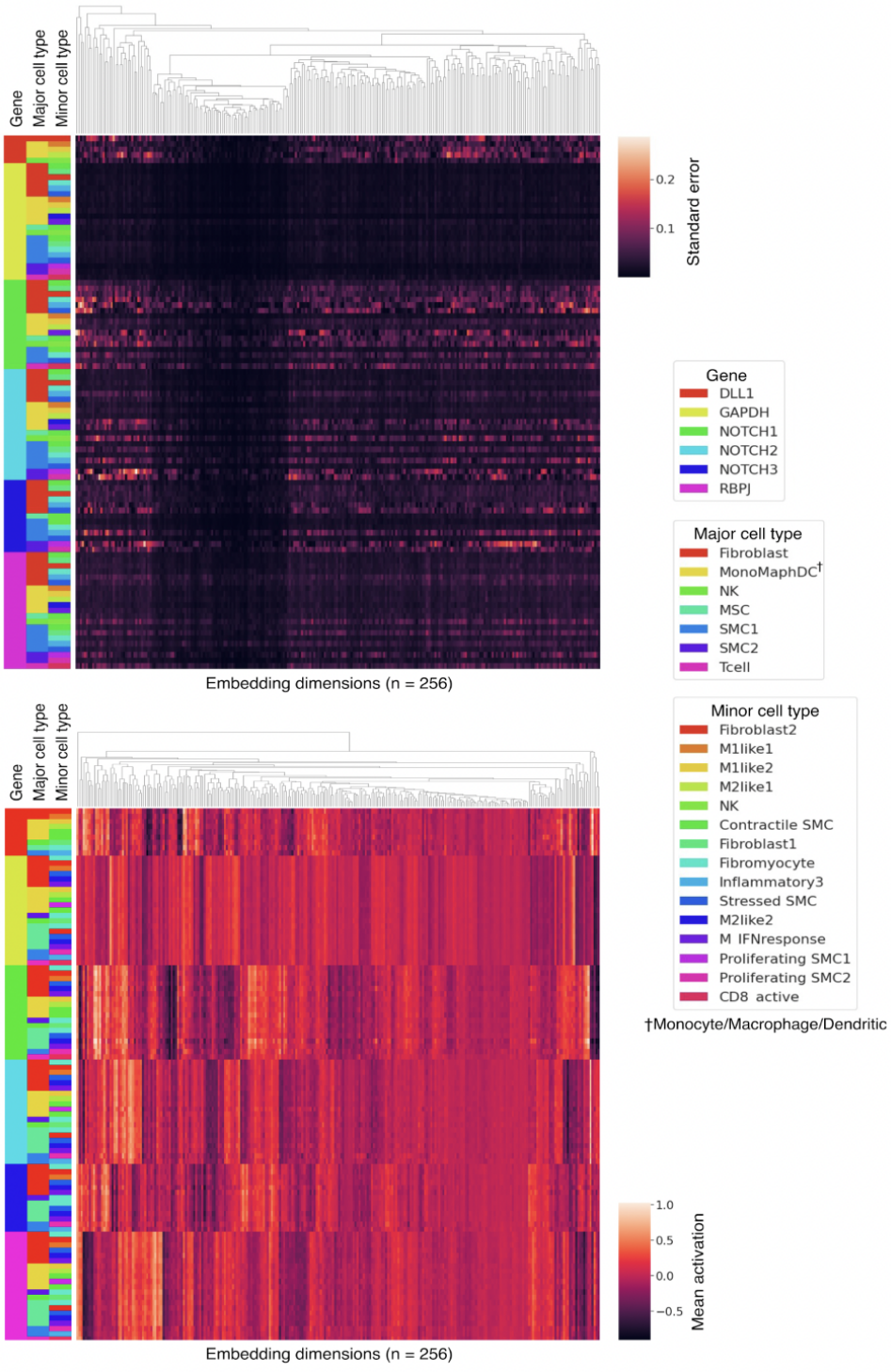
We also added an example plot of embedding values for the model fine-tuned to distinguish 16 lung cell types (below and Extended Data Fig. 6g). In this fine-tuned model, the embedding dimensions encode the identity of the lung cell types the model was fine-tuned to distinguish.

Finally, we also plotted each layer's embedding of cells within iPSC to cardiomyocyte differentiation in Extended Data Fig. 4c to visualize how each layer progressed towards the final classification by the model. While the initial embedding does not separate the cells by cell type

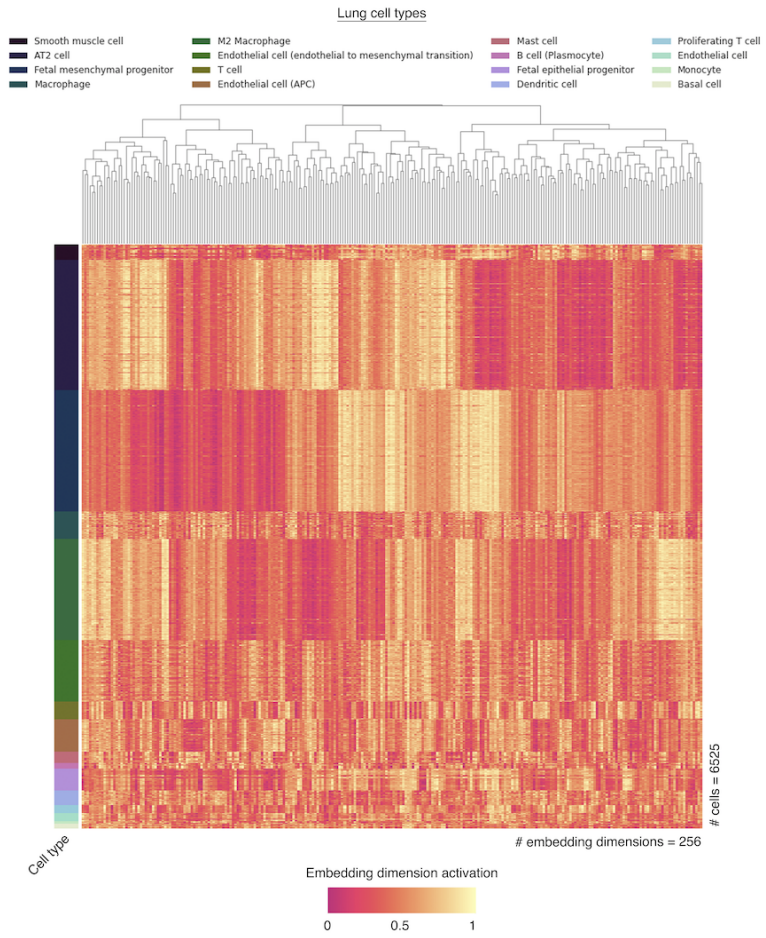
class, each successive layer progressively separates the classes until they are highly separable. Thus, each layer's embedding contributes to further separating the classes based on the attended features in that layer.

Manuscript changes:

- Addition of bottom plot in Extended Data Fig. 3 (below):



- Addition of Extended Data Fig. 6g (below):



7. Code is not available. To assess the model in this work, in terms of reproducibility and performance, it is essential to access to the source code and the scripts used to generate figures. The description of model architecture and concepts in the manuscript is not sufficient to fully prove if the model is working, and if the figures are reliable and reproducible. Overall the model architecture is relatively simple with 6 transformer units, each containing a self attention layer (4 attention heads) and a feed forward fully connected network (512 neurons) to get embeddings in 256 dimensions. For benchmaring with other methods, this work only compared to simple classifiers like random forest, support vector machine and logistic regression, but not with other deep learning methods. The codes for comparisons need to be available to assess the reliability and reproducibility.

Authors' response: We thank the reviewer for this important point and provide the Geneformer model and related code at https://huggingface.co/ctheodoris/Geneformer and Genecorpus-30M pretraining corpus at https://huggingface.co/datasets/ctheodoris/Genecorpus-30M on the Huggingface Hub (a GitHub-based platform for sharing machine learning models and related datasets).

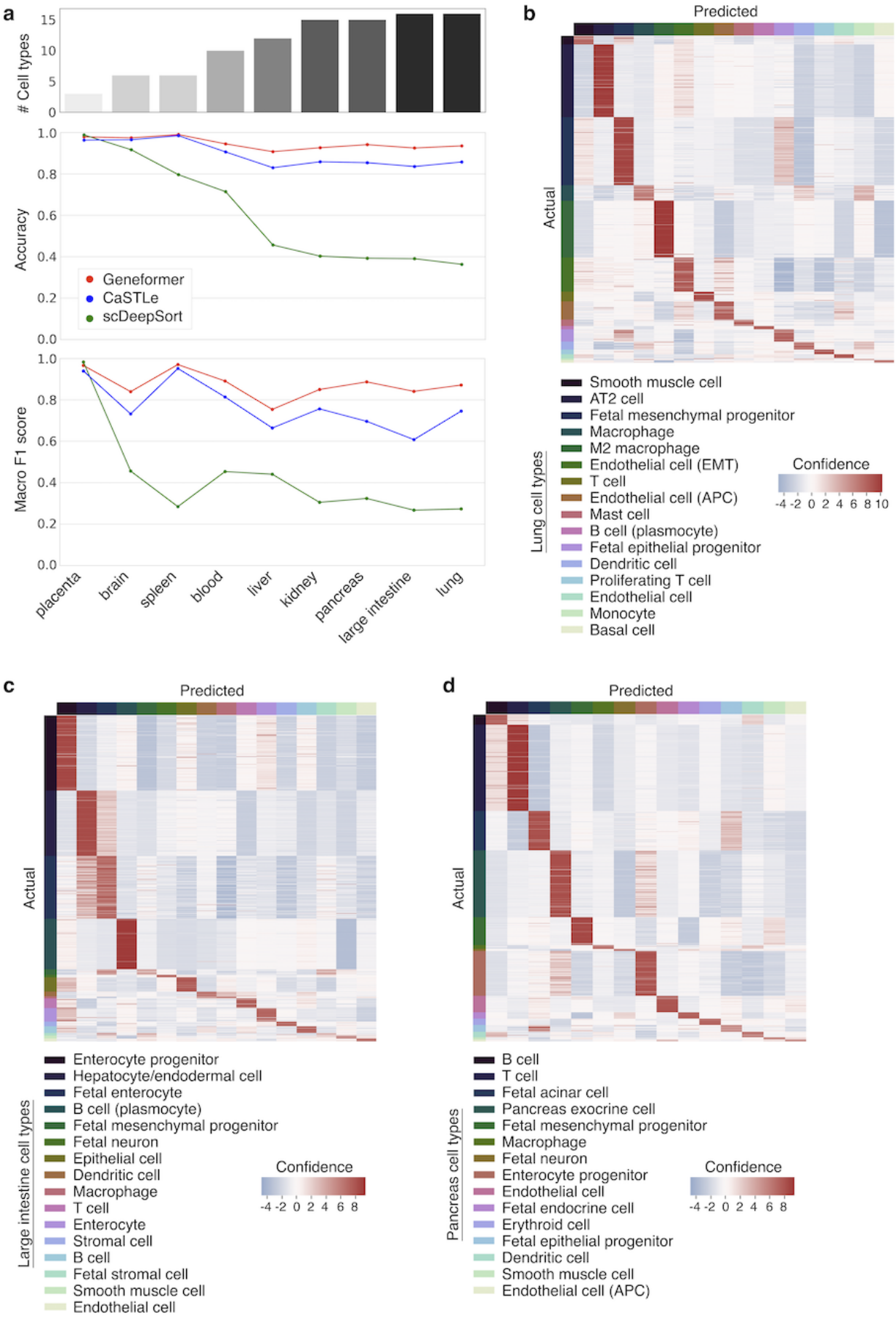
Benchmarking scripts are also available on the Huggingface Hub at https://huggingface.co/ctheodoris/Geneformer. We also would like to note that for benchmarking, in addition to the classifiers mentioned by the reviewer, we did also compare to attention-based deep learning architectures with varying number of layers with retained width-to-depth aspect ratios that have not been pretrained as well as attention-based deep learning models of the same architecture as Geneformer that have been pretrained with variable amounts and diversity of pretraining data (Fig. 2a-b). In response to reviewer #2 comment #B2, we also added comparison to another deep learning method for cell type annotation (Extended Data Fig. 5-6).

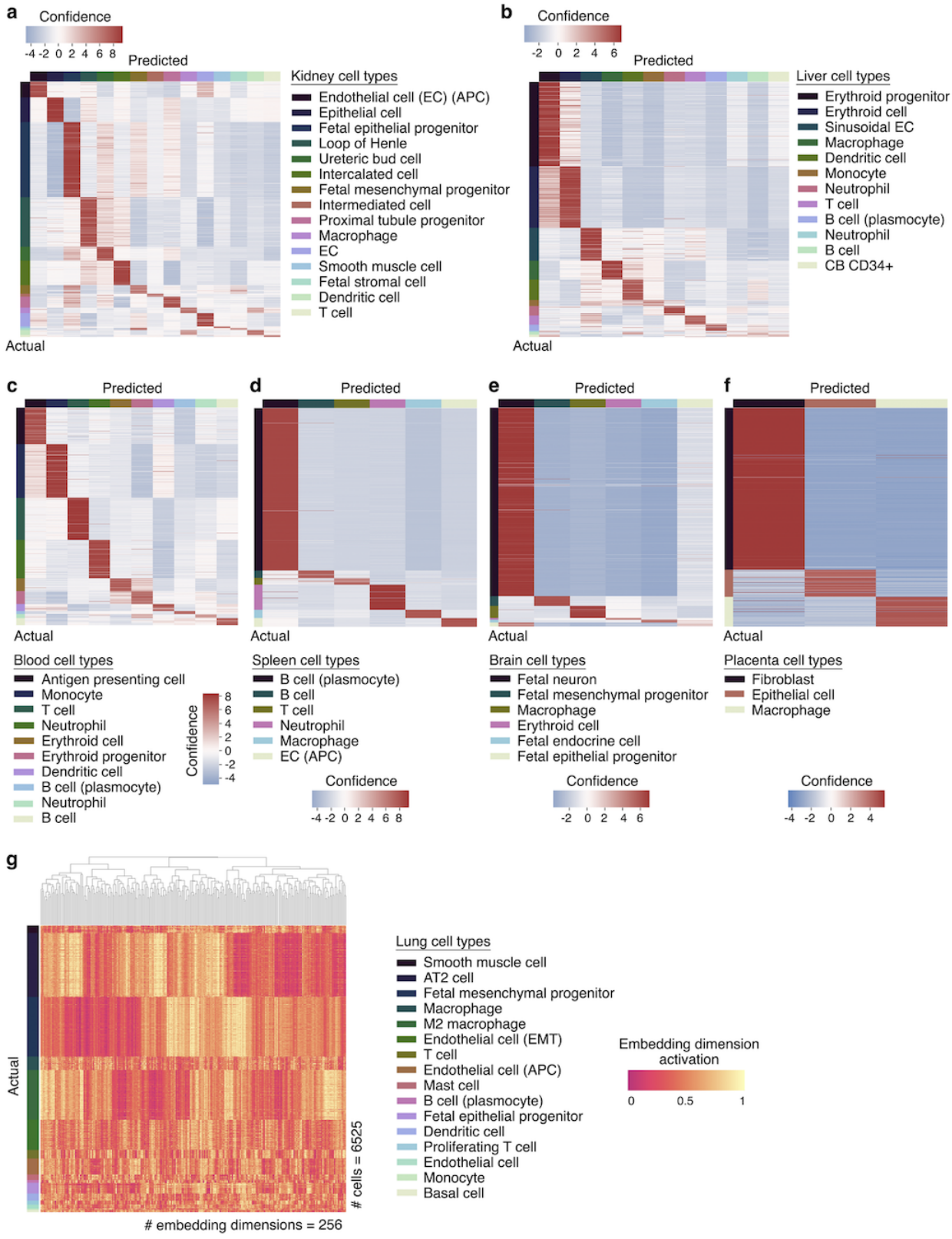
Manuscript changes:

- Updated code availability statement as above.

- Addition of Extended Data Fig. 5-6 (below) and relevant text:

“Although Geneformer is most focused on understanding network dynamics rather than cell-level annotations, we further investigated Geneformer’s performance in cell type annotation given it is a common application for previously published models. We compared Geneformer to alternative XGBoost and deep neural network-based models. These methods train a new model from scratch for each separate tissue using the same supervised learning objective as is used for the final cell type predictions in that specific tissue. Therefore, these approaches do not take advantage of the large amounts of data available more broadly that are not specifically labeled for that task. In contrast, Geneformer learns from large-scale unlabeled data during the self-supervised pretraining using a generalizable learning objective to gain fundamental knowledge that can then be transferred to a multitude of new and diverse fine-tuning tasks. Compared to these alternative methods, Geneformer boosted cell type predictions in a variety of tissues, with the gap in performance by accuracy and macro F1 score increasing as the number of cell type classes increased, indicating that Geneformer was robust in even increasingly complex multiclass prediction applications (Extended Data Fig. 5-6). “


a
Cell types
Accuracy
Macro F1 score
Geneformer
CaSTLe
scDeepSort
placenta
brain
spleen
blood
liver
kidney
pancreas
large intestine
lung
b
Predicted
Actual
Lung cell types
Smooth muscle cell
AT2 cell
Fetal mesenchymal progenitor
Macrophage
M2 macrophage
Endothelial cell (EMT)
T cell
Endothelial cell (APC)
Mast cell
B cell (plasmocyte)
Fetal epithelial progenitor
Dendritic cell
Proliferating T cell
Endothelial cell
Monocyte
Basal cell
Confidence
c
Predicted
Actual
Large intestine cell types
Enterocyte progenitor
Hepatocyte/endodermal cell
Fetal enterocyte
B cell (plasmocyte)
Fetal mesenchymal progenitor
Fetal neuron
Epithelial cell
Dendritic cell
Macrophage
T cell
Enterocyte
Stromal cell
B cell
Fetal stromal cell
Smooth muscle cell
Endothelial cell
Confidence
d
Predicted
Actual
Pancreas cell types
B cell
T cell
Fetal acinar cell
Pancreas exocrine cell
Fetal mesenchymal progenitor
Macrophage
Fetal neuron
Enterocyte progenitor
Endothelial cell
Fetal endocrine cell
Erythroid cell
Fetal epithelial progenitor
Dendritic cell
Smooth muscle cell
Endothelial cell (APC)
Confidence


a
Confidence
-4 -2 0 2 4 6 8
Predicted
Actual
Kidney cell types
Endothelial cell (EC) (APC)
Epithelial cell
Fetal epithelial progenitor
Loop of Henle
Ureteric bud cell
Intercalated cell
Fetal mesenchymal progenitor
Intermediated cell
Proximal tubule progenitor
Macrophage
EC
Smooth muscle cell
Fetal stromal cell
Dendritic cell
T cell
b
Confidence
-2 0 2 4 6
Predicted
Actual
Liver cell types
Erythroid progenitor
Erythroid cell
Sinusoidal EC
Macrophage
Dendritic cell
Monocyte
Neutrophil
T cell
B cell (plasmocyte)
Neutrophil
B cell
CB CD34+
c
Predicted
Actual
Blood cell types
Antigen presenting cell
Monocyte
T cell
Neutrophil
Erythroid cell
Erythroid progenitor
Dendritic cell
B cell (plasmocyte)
Neutrophil
B cell
Confidence
8 6 4 2 0 -2 -4
d
Predicted
Actual
Spleen cell types
B cell (plasmocyte)
B cell
T cell
Neutrophil
Macrophage
EC (APC)
Confidence
-4 -2 0 2 4 6 8
e
Predicted
Actual
Brain cell types
Fetal neuron
Fetal mesenchymal progenitor
Macrophage
Erythroid cell
Fetal endocrine cell
Fetal epithelial progenitor
Confidence
-2 0 2 4 6
f
Predicted
Actual
Placenta cell types
Fibroblast
Epithelial cell
Macrophage
Confidence
-4 -2 0 2 4
g
Actual
cells = 6525
embedding dimensions = 256
Lung cell types
Smooth muscle cell
AT2 cell
Fetal mesenchymal progenitor
Macrophage
M2 macrophage
Endothelial cell (EMT)
T cell
Endothelial cell (APC)
Mast cell
B cell (plasmocyte)
Fetal epithelial progenitor
Dendritic cell
Proliferating T cell
Endothelial cell
Monocyte
Basal cell
Embedding dimension activation
0 0.5 1

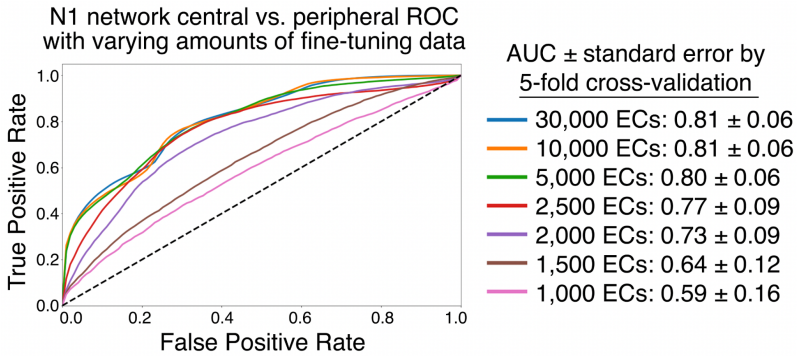
Minor limitations

1. Practicality of the tool for prediction

It is important for readers to know practicality of this tool when applying their given input of interest. Importantly, the paper did not indicate how much input data are needed to yield confident predictions. The authors used limited data for fine-tuning (e.g. 30k for gene network hierarchy, which is certainly not small in many cases) are sufficient but it's not clear how this variable influences the outcome. Are there any thresholds (i.e. minimum number required?) or totally data-dependent? I assume there should be a certain range of cell numbers needed to adjust the pre-defined model assuming acceptable sequencing depth. This can be an important question to be answered as often not such cell numbers are available for a cell type of interest.

Authors' response: We thank the reviewer for this question. As the reviewer suggested, the minimum number of cells required for fine-tuning is dependent on the application (as noted by the varied amount of task-specific data for each of the applications demonstrated in this work) as well as relevance of the fine-tuning data to the task at hand. We further characterized this within the N1 network hierarchy application by progressively reducing the number of endothelial cells (ECs) used for the fine-tuning (below and Fig. 4c). As few as 5,000 ECs yielded similar results to using the total 30,000 ECs available within the Heart Atlas[18], but predictive potential decreased with fewer numbers of cells used. However, when using ECs from a dataset encompassing ECs from patients with healthy aortas or aortic aneurysms[19], predictive potential from fine-tuning with only 884 ECs (Fig. 4d) was similar to the results of using 2,000 normal ECs from the Heart Atlas. Thus, using data that is most relevant to the downstream task, especially if encompassing relevant perturbations or disease states, will likely yield improved predictive potential with fewer cells.

Manuscript changes: Addition of Fig. 4c (below) and relevant text:



"To investigate the threshold for minimal data needed for fine-tuning, we fine-tuning the pretrained Geneformer with progressively smaller numbers of normal ECs from the Heart Atlas to distinguish central versus peripheral factors within the N1-dependent gene network. We found that nearly equivalent predictive potential was retained even when reducing the fine-tuning data to only 5,000 ECs (Fig. 4c). Then, to determine whether Geneformer could generate meaningful predictions using an even more miniscule number of fine-tuning training examples when the

task-specific data was more relevant to the learning objective, we fine-tuned the pretrained Geneformer using only 884 ECs from healthy versus dilated aortas. Interestingly, Geneformer was able to distinguish central versus peripheral factors in the N1-dependent network with fine-tuning on this very minimal data to a better degree than the alternative methods' predictions trained on the larger dataset of ~30,000 ECs, demonstrating the strength of pretraining in enabling predictions from increasingly limited data (Fig. 4d, Extended Data Fig. 9b). More than twice as many general cardiac ECs were needed to gain similar predictive potential as was possible from fine-tuning with the more relevant data from healthy versus dilated aortas, suggesting that the minimum amount of fine-tuning data needed is dependent on both the specific application and relevance of the data to that task."

2. Unclear justification of the model selection
Authors defined the neural network with specified hyperparameters (e.g. 5 layers, 256 dimensions etc.) without clear justification. While these parameters were presumably selected for their optimal performance, it remains unclear to readers how this was done relevant details were not included. It would be ideal to include such details so that the model selection was done in a reproducible way. It is understandable that this particular ML model can often appear as a "blackbox" due to low-interpretability. Despite that, I think it's critical to justify selection of a given model from computational perspective – how the training was performed (e.g. separation of datasets for training, validation and optimization).

Authors' response: We thank the reviewer for this question. As a general principle, the goal is to train the deepest attention-based model architecture for which there is sufficient data to train. This approach has been established to yield the greatest predictive potential in other informational fields including NLU, computer vision, and mathematical problem-solving[4]. The width-to-depth aspect ratio was chosen based on ratios found to be effective in analogous attention-based models[5], and this ratio was accordingly scaled to the maximum depth achievable based on our pretraining data quantity and available GPU resources. We also maximized the amount of context considered by the model with full attention, which is considerably larger than standard language models and contributes to the compute required for pretraining given the quadratic memory and time complexity of full attention. We implemented recent advances in distributed GPU training[6,7] to allow efficient pretraining on the large-scale corpus within the available GPU resources.

In the future though, larger pretraining corpuses and further advances in GPU hardware and distribution algorithms will likely allow training of deeper Geneformer models, which may open opportunities to achieve meaningful predictions in even more elusive tasks with increasingly limited task-specific data. Additionally, the input size of Geneformer fully represented 93% of the rank value encodings in Genecorpus-30M, so based on the number of genes standardly detected in each cell using the current scRNA-seq technology, we were able to take advantage of the predictive benefits of full attention (rather than sparse attention) despite the quadratic memory and time complexity. However, future advancements in scRNA-seq technology or reduced sequencing costs may allow detection of larger numbers of genes per cell. At that time,

consideration of the greater context size may offer predictive benefits that outweigh the benefits of applying full attention, so future work may design a model to leverage sparse attention to allow consideration of greater context size without the limitations of the quadratic memory and time complexity of full attention.

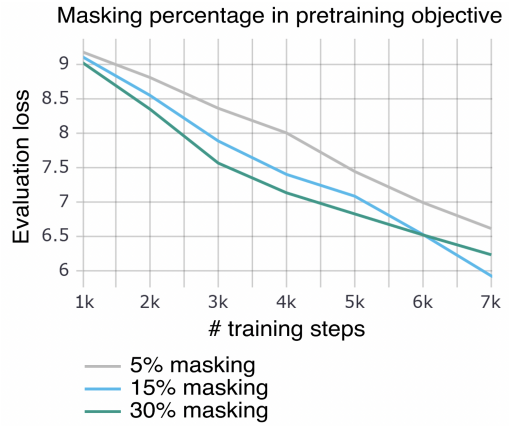
Furthermore, future advances in GPU hardware and distribution algorithms may allow further optimization of model architectures and training hyperparameters, extensive optimization of which at the pretraining stage is currently precluded by the large amount of resources needed for pretraining and the fact that optimization with subsampled datasets is known not to generalize to larger datasets due to the scaling laws of these models[4]. Of note, training hyperparameter optimization at the fine-tuning stage is computationally feasible and highly recommended as these hyperparameters will be data- and application-specific. Training hyperparameters were tuned for the cardiomyopathy disease modeling fine-tuning application as described in the Extended Methods ("Disease Modeling" section).

Regarding the masking percentage, 15% was chosen because it is the standard percentage used for masked learning objectives in other informational fields[5]. However, we repeated pretraining on a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for evaluation, and found that 15% masking had marginally lower evaluation loss compared to 5% or 30% masking (below and Extended Data Fig. 1e).

Regarding the separation of datasets for optimization, training, and validation, all fine-tuning applications aside from the disease modeling were fine-tuned with standardized hyperparameters without optimization to examine the baseline efficacy of the pretrained Geneformer in boosting predictive potential of downstream applications. It should be noted that hyperparameter tuning for deep learning applications generally significantly enhances learning and so it is likely that the maximum predictive potential of Geneformer in these downstream applications is significantly underestimated. The predictive potential of these fine-tuned models was quantified by 5-fold cross-validation. For the disease modeling application, we optimized hyperparameters based on performance on a held-out portion of the training data and then fine-tuned the model using all of the training data with the optimal hyperparameters. The predictive potential of the trained model was then quantified on completely separate out of sample data not used for training or hyperparameter optimization.

Manuscript changes:

- Addition of Extended Data Fig. 1e (below):



- Addition of text regarding parameter choice:

"Depth was chosen based on the maximum depth for which there was sufficient data to pretrain as it has been established that this approach yields the greatest predictive potential in other informational fields including NLU, computer vision, and mathematical problem-solving. Additionally, we maximized the amount of context (input size) considered by the model with full attention based on the number of genes standardly detected in each cell within the pretraining corpus."

3. Genes affected by in-silico deletion
If authors could identify downstream genes which were affected by the in-silico deletions performed in Fig 2d and characterise them using curated databases such as KEGG or GO, it would help provide more context for the mechanism of the in-silico deletions and their downstream effects.

Authors' response: We thank the reviewer for this suggestion. Overall, genes whose deletion was predicted to have the most deleterious effect on cardiomyocytes were significantly enriched for human phenotypes including cardiomyopathy and abnormal myocardial morphology and processes including mitochondrial organization, which is critical for providing the energy needed for cardiomyocyte contraction (Extended Data Table 3).

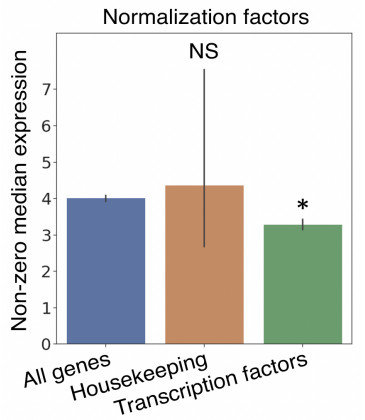
Manuscript changes: Addition of Extended Data Table 3 (gene set enrichments of genes whose in silico deletion is predicted to be deleterious in fetal cardiomyocytes).

3. Normalisation method
Authors ranked genes within each cell by normalising by that gene's expression across the entire Genecorpus. Is it possible that this strategy may prioritise genes that are expressed highly in a cell-type specific manner, over genes expressed lowly, such as transcription factors?

Authors' response: We thank the reviewer for this comment. We quantified the normalization factors for housekeeping genes and transcription factors compared to all genes to better

characterize the effects of the normalization (below and in Extended Data Fig. 1c). Transcription factors are normalized by a statistically significantly lower factor (resulting in higher prioritization in the rank value encoding) compared to all genes. Housekeeping genes on average show a trend of a higher normalization factor (resulting in deprioritization in the rank value encoding) compared to all genes. Because we utilize the non-zero median value of expression for normalization, genes that are expressed highly in a cell-type specific manner will still be normalized by a higher factor (and thus deprioritized) compared to transcription factors that are lowly expressed.

Manuscript changes: Addition of Extended Data Fig. 1c (below):



4. In silico over-expression of genes
Aside from the OKSM factors, the paper only utilises in silico deletion. Only utilising deletion could miss important biological aspects, particularly in disease and therapeutic contexts. Can in-silico overexpression be applied to the cardiomyopathy dataset?

Authors' response: We appreciate the reviewer's suggestion. We initially focused on in silico deletion because inhibition is generally a more straightforward therapeutic strategy. However, we now also include in silico activation analysis for the cardiomyopathy dataset in Extended Data Tables 3, 6, 8, 10, 11, and 13. Genes whose in silico activation in cardiomyocytes from hypertrophic cardiomyopathy patients significantly shifted the embeddings towards the non-failing state were enriched for genes involved in cardiac muscle cell development, Z discs, contractile fibers, and oxidative phosphorylation. Candidate therapeutic targets for activation included VEGFA, which previously has been shown to protect cardiac function in hypertrophic hearts (Hu et al, *Journal of Cardiothoracic Surgery* 2011), NEXN, whose overexpression has been shown to attenuate cardiac hypertrophy (Friedrich et al, *Basic Res Cardiol* 2014), and NEXN, damaging variants in which cause hypertrophic cardiomyopathy (Wang et al, *Am J Hum Genet* 2010).

Manuscript changes: Addition of in silico activation analysis to Extended Data Tables 3, 6, 8, 10, 11, and 13.

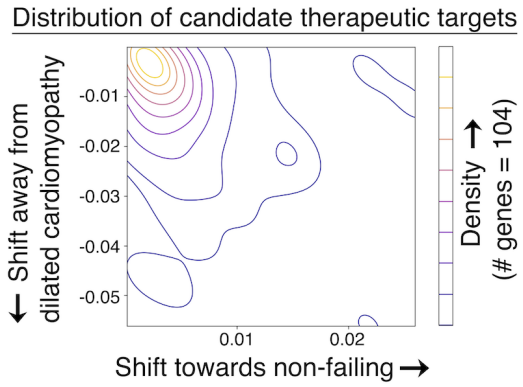
5. Interpretation of significance
Some of result data show dubious significance, for example in Fig 6d. These targets are statistically marginally enriched. As p-value is largely affected by the sample size (assuming n=104), would fold-change be a useful alternative?

Authors' response: While we agree with the reviewer that the p-value is definitely affected by sample size (n=104 as the reviewer indicated), we still feel that p-value adjusted for multiple hypothesis testing is the most appropriate way to represent the data in Extended Data Fig. 10d (previously Fig. 6d) as it accounts for variance and even if other pathways showed larger fold-change, if they were not statistically significant it would be dubious to present them.

6. Absence of scale bar
Figure 6e contour plot does not include the scale bar for gene number while the interpretation of the result is not affected by this information.

Authors' response: We added a color scale bar for the kernel density estimate plot in Fig. 6e. The color indicates relative density (points per unit square), and the total number of genes (104) are now indicated.

Manuscript changes: Addition of color scale bar to Fig. 6e (below):



References:

1. Ang, Y.-S. *et al.* Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis. *Cell* **167**, 1734–1749.e22 (2016).
2. Shao, X. *et al.* scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res.* **49**, e122 (2021).
3. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* **13**, e0205499 (2018).
4. Henighan, T. *et al.* Scaling Laws for Autoregressive Generative Modeling. *arXiv [cs.LG]*

(2020).
5. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).
6. Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. *arXiv [cs.LG]* (2019).
7. Ren, J. *et al.* ZeRO-Offload: Democratizing Billion-Scale Model Training. *arXiv [cs.DC]* (2021).
8. Vaswani, A. *et al.* Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
9. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.* **366**, 619–628 (2012).
10. Hinson, J. T. *et al.* HEART DISEASE. Titin mutations in iPS cells define sarcomere insufficiency as a cause of dilated cardiomyopathy. *Science* **349**, 982–986 (2015).
11. Guo, D. *et al.* iMyoblasts for ex vivo and in vivo investigations of human myogenesis and disease modeling. *Elife* **11**, (2022).
12. Kathiriya, I. S. *et al.* Modeling Human TBX5 Haploinsufficiency Predicts Regulatory Networks for Congenital Heart Disease. *Dev. Cell* **56**, 292–309.e9 (2021).
13. Verma, P. & Berger, J. Audio Transformers:Transformer Architectures For Large Scale Audio Understanding. Adieu Convolutions. *arXiv [cs.SD]* (2021).
14. Selva, J. *et al.* Video Transformers: A Survey. *arXiv [cs.CV]* (2022).
15. Selewa, A. *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Sci. Rep.* **10**, 1535 (2020).
16. A Mathematical Framework for Transformer Circuits. https://transformer-circuits.pub/2021/framework/index.html.
17. In-context learning and induction heads. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.
18. Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466–472 (2020).
19. Li, Y. *et al.* Single-Cell Transcriptome Analysis Reveals Dynamic Cell Populations and Differential Gene Expression Patterns in Control and Aneurysmal Human Aortic Tissue. *Circulation* **142**, 1374–1388 (2020).

Reviewer Reports on the First Revision:

Referee expertise:

Referee #1: systems biology

Referee #2: network analysis

Referee #3: cardiac/computational biology

Referees' comments:

Referee #1 (Remarks to the Author):

Two of my major concerns raised in the previous review have not been satisfactorily addressed., Addressing these concerns fully would be necessary to convince the readers of the method's applicability.

1) Normalization method: In the previous review, I mentioned that the normalization method may lead to overestimating the value of genes whose median expression is extremely low (by normalizing them by a very low number). A similar concern was raised by the other reviewers as well.

In their response, the authors present in Extended Data Fig. 1c the average normalization factors for different groups of genes. While the theoretical motivation of the normalization is explained, the average analysis does not address the potential issue of individual genes that were normalized by a very small number. This may lead to extreme normalized values for such individual genes. The data of low expressed genes are highly affected by stochastic factors as well as sequencing depth, so this may have a large, and unpredictable effect on the results when applied to new datasets.

2) In silico perturbations:

2.1) The authors state that the results of deletion of GATA4 and TBX5 suggest that in silico perturbation can be applied to model gene network connections. However, it is not clear how many tests on how many genes have been performed. The proposed theory that genes most affected by a given gene's perturbation would be enriched for that gene's biological downstream targets, should be supported by systematic tests using many perturbations and measuring whether they have statistically significant results.

2.2) Furthermore, have the results of the in silico deletion of both GATA4 and TBX5 in combination (Fig. 5b) been validated with real data?

Referee #2 (Remarks to the Author):

The authors have addressed all my comments successfully. I have no further comments.

Referee #3 (Remarks to the Author):

Reviewer comments – Theodoris

We thank the authors of this manuscript for completing additional work to address concerns around this study. The additional work and inclusions in the manuscript have largely addressed our concerns and strengthened the manuscript as a whole. However, there are still several minor points remaining.

The authors have functionally validate the role of TEAD4, a dosage-sensitive gene predicted by Geneformer, in cardiac development. This is consistent with a role for the HIPPO pathway in cardiovascular differentiation (see reference 1 below). The validation of a single target is interesting and valuable, however, we would like to propose another reasonably straightforward computational analysis that would provide a powerful link to human biology and increase the value of the method significantly. To strengthen the biological power of Geneformer's predictions of context-specific dosage sensitive genes, we suggest referring to statistical genetics data of cardiac MRI traits (see reference 2) to determine if genes predicted to be dosage-sensitive in the heart by Geneformer are present among genes significantly associated with variation in human cardiac MRI traits. This analysis would test the hypothesis that Geneformer predicted genes significantly influence physiological parameters of human heart structure and function. Collectively, this would not only account for both known and novel genes from Geneformer, but it would also provide validation across a larger number of gene candidates and provide a link with human health and disease.

The authors further demonstrated that Geneformer distinguishes the effects of in silico GATA4 knockout on GATA4 targets from targets of other transcription factors using direct targets of NKX2-5 as a control. The use of NKX2-5 targets as a control set is questionable given the highly related nature of these transcription factors. We suggest the use of direct targets of an unrelated transcription factor. Furthermore, a more interesting and convincing analysis regarding the context-awareness of Geneformer would be to test whether Geneformer can capture and predict the biology of heterotypic interactions between multiple transcription factors, such as those described by Luna-Zurita et al. (reference 3)

The authors completed further work to effectively demonstrate Geneformer's capability to model in silico reprogramming. In the response to reviewer response, the authors state that "The end state cells define the coordinates of the end point to specify the goal direction of the embedding vector shift. However, the embeddings of genes in the end state will always be more different than the embeddings of genes in the perturbed cell because the context in the end cell is the biological end state of the perturbation whereas the perturbed cell is encoded as only the first step in the

downstream changes that would biologically occur in response to the perturbation." Can the authors properly test this claim by testing Geneformer against data captured across a time-course of reprogramming (e.g. Zhou et al. reference 4) Doing so would help clarify whether Geneformer is capturing the changes associated with the early (and not all) stages of reprogramming and help users understand both the strengths and limitations of the method.

Lastly, we suggest that the authors compare the gene perturbations predicted by Geneformer in response to in silico deletion of TBX5 or GATA4 to gene expression changes in a RNA sequencing dataset of TBX5 or GATA4 knockout. For instance, in GATA4 knockout some direct targets are predicted to change more than others. We believe comparing changes predicted by Geneformer to a RNA sequencing dataset of the knockout will validate whether Geneformer is faithfully modelling gene knockout.

1. Nelakanti,R. V and Xiao,A.Z. (2020) A New Link to Primate Heart Development. Dev. Cell, 54, 685–686.
2. Pirruccello,J.P., Bick,A., Wang,M., Chaffin,M., Friedman,S., Yao,J., Guo,X., Venkatesh,B.A., Taylor,K.D., Post,W.S., et al. (2020) Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat. Commun., 11, 2254.
3. Luna-Zurita,L., Stirnimann,C.U., Glatt,S., Kaynak,B.L., Thomas,S., Baudin,F., Samee,M.A.H., He,D., Small,E.M., Mileikovsky,M., et al. (2016) Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. Cell, 164, 999–1014.
4. Zhou,Y., Liu,Z., Welch,J.D., Gao,X., Wang,L., Garbutt,T., Keepers,B., Ma,H., Prins,J.F., Shen,W., et al. (2019) Single-Cell Transcriptomic Analyses of Cell Fate Transitions during Human Cardiac Reprogramming. Cell Stem Cell, 25, 149-164.e9.

Author Rebuttals to First Revision:

The reviewers' comments are shown in **black** and our responses are in **blue.**

Referee One

Two of my major concerns raised in the previous review have not been satisfactorily addressed. Addressing these concerns fully would be necessary to convince the readers of the method's applicability.

1) Normalization method: In the previous review, I mentioned that the normalization method may lead to overestimating the value of genes whose median expression is extremely low (by normalizing them by a very low number). A similar concern was raised by the other reviewers as well.

In their response, the authors present in Extended Data Fig. 1c the average normalization factors for different groups of genes. While the theoretical motivation of the normalization is explained, the average analysis does not address the potential issue of individual genes that were normalized by a very small number. This may lead to extreme normalized values for such individual genes. The data of low expressed genes are highly affected by stochastic factors as well as sequencing depth, so this may have a large, and unpredictable effect on the results when applied to new datasets.

We thank the reviewer for this point. We fully agree that the normalization must be consistent between all datasets to avoid unpredictable effects. In the provided code, the tokenizer that converts the initial count matrix to the rank value encoding includes this normalization step. As such, every future dataset used with this model will have each gene normalized by the equivalent factor as it was for the pretraining of the model to ensure consistency. We apply the model to multiple new datasets throughout the manuscript to demonstrate applicability to new data.

To ensure this process is clear to users, we added the following statement to the methods:

"The provided tokenizer code includes this normalization step and should be used for tokenizing new datasets presented to Geneformer to ensure consistency of the normalization factor used for each gene."

2) In silico perturbations:

2.1) The authors state that the results of deletion of GATA4 and TBX5 suggest that in silico perturbation can be applied to model gene network connections. However, it is not clear how many tests on how many genes have been performed. The proposed theory that genes most affected by a given gene's perturbation would be enriched for that gene's biological downstream targets, should be supported by systematic tests using many perturbations and measuring whether they have statistically significant results.

We agree with the reviewer's point. However, we also feel it is important that the model be compared to a ground truth that is known with strong confidence. For this reason we chose to compare against datasets that had very reliable paired ChIP-seq and RNA-seq defining the direct targets of the tested transcription factors. The tested comparisons are included in the manuscript and showed statistically significant results.

In theory, Perturb-seq data would be an ideal method for testing our approach; however, current Perturb-seq datasets are limited. The available studies we evaluated had minimal perturbations with significant effects on the intended target or were based on very few cells for each target leading to very few differentially expressed targets. We are hopeful that in the future the technology will be more robust such that we can fine-tune Geneformer specifically for the task of predicting perturbations by training on data from such systematic perturbation studies and applying the model to predict unseen perturbations. We are also hopeful that in the future there will be large-scale datasets of transcription factor ChIP-seq matched with these perturbations in the same cell types so that the direct targets can be confidently defined. However, currently, robust datasets of paired ChIP-seq and perturbation RNA-seq are limited.

2.2) Furthermore, have the results of the in silico deletion of both GATA4 and TBX5 in combination (Fig. 5b) been validated with real data?

We thank the author for requesting this clarification. Yes, the co-bound targets are validated by real data from ChIP-seq experiments from Ang et al. *Cell* 2016. We added the citation to the text where this is discussed to clarify this point.

Referee Two

The authors have addressed all my comments successfully. I have no further comments.

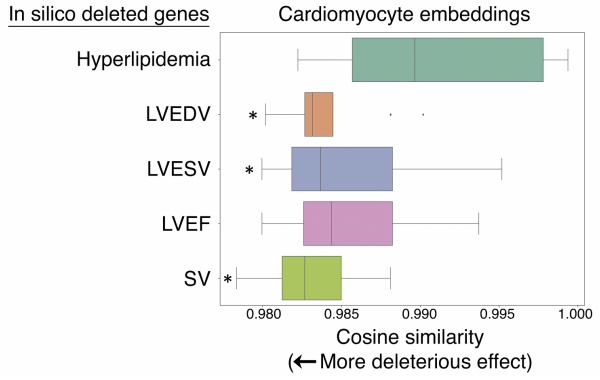
Referee Three

We thank the authors of this manuscript for completing additional work to address concerns around this study. The additional work and inclusions in the manuscript have largely addressed our concerns and strengthened the manuscript as a whole. However, there are still several minor points remaining.

The authors have functionally validated the role of TEAD4, a dosage-sensitive gene predicted by Geneformer, in cardiac development. This is consistent with a role for the HIPPO pathway in cardiovascular differentiation (see reference 1 below). The validation of a single target is interesting and valuable, however, we would like to propose another reasonably straightforward computational analysis that would provide a powerful link to human biology and increase the value of the method significantly. To strengthen the biological power of Geneformer's predictions of context-specific dosage sensitive genes, we suggest referring to statistical genetics data of cardiac MRI traits (see reference 2) to determine if genes predicted to be dosage-sensitive in the heart by Geneformer are present among genes significantly associated with variation in human cardiac MRI traits. This analysis would test the hypothesis that Geneformer predicted genes significantly influence physiological parameters of human heart structure and function. Collectively, this would not only account for both known and novel genes from Geneformer, but it would also provide validation across a larger number of gene candidates and provide a link with human health and disease.

We appreciate the reviewer's suggestion. In silico deletion of genes linked by the genome-wide association study (GWAS) referenced by the reviewer to cardiac magnetic resonance imaging (MRI) traits of LVEDV, LVESV, and SV had a significantly larger effect compared to the control set.

Addition of Extended Data Fig. 7b:

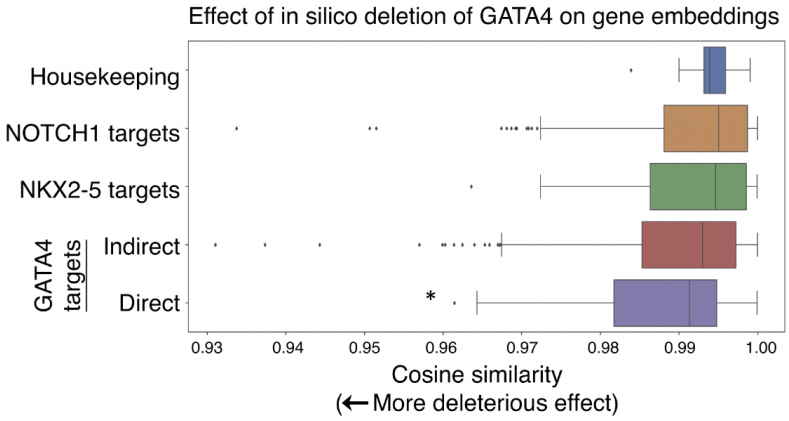


Addition to text:

"In silico deletion of genes linked by a prior genome-wide association study (GWAS) to cardiac magnetic resonance imaging (MRI) traits relevant to cardiac disease also had a larger effect compared to the control set (Extended Data Fig. 7b)."

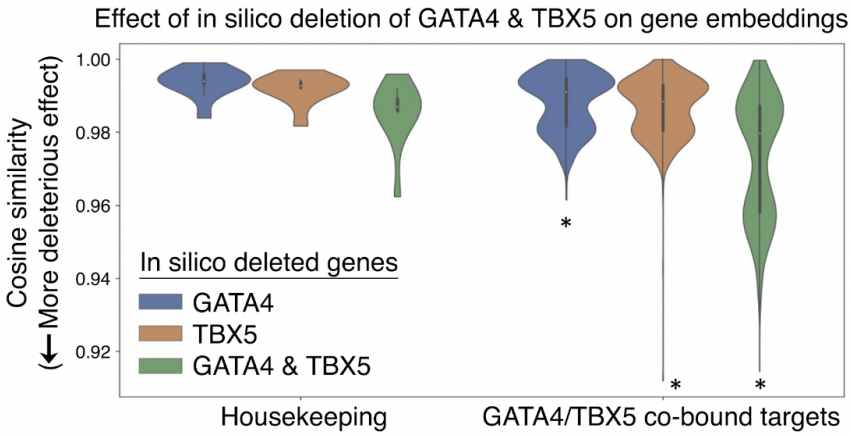
The authors further demonstrated that Geneformer distinguishes the effects of in silico GATA4 knockout on GATA4 targets from targets of other transcription factors using direct targets of NKX2-5 as a control. The use of NKX2-5 targets as a control set is questionable given the highly related nature of these transcription factors. We suggest the use of direct targets of an unrelated transcription factor.

We appreciate the reviewer's comments. We added another control of direct targets of NOTCH1 to Fig. 5a to address this point:



Furthermore, a more interesting and convincing analysis regarding the context-awareness of Geneformer would be to test whether Geneformer can capture and predict the biology of heterotypic interactions between multiple transcription factors, such as those described by Luna-Zurita et al. (reference 3)

We agree with the reviewer that predicting interactions between transcription factors is of great interest. Luna-Zurita et al. *Cell* 2016 investigated interactions between cardiomyocyte transcription factors including GATA4 and TBX5. We address this within the analysis in Fig. 5b, which demonstrates that Geneformer models synergistic interactions between these two factors on their ChIP-seq-defined co-bound targets (defined in a contemporaneous paper to the one the reviewer suggested, Ang et al. *Cell* 2016):



The authors completed further work to effectively demonstrate Geneformer's capability to model in silico reprogramming. In the response to reviewer response, the authors state that "The end state cells define the coordinates of the end point to specify the goal direction of the embedding vector shift. However, the embeddings of genes in the end state will always be more different than the embeddings of genes in the perturbed cell because the context in the end cell is the biological end state of the perturbation whereas the perturbed cell is encoded as only the first step in the downstream changes that would biologically occur in response to the perturbation." Can the authors properly test this claim by testing Geneformer against data captured across a time-course of reprogramming (e.g. Zhou et al. reference 4) Doing so would help clarify whether Geneformer is capturing the changes associated with the early (and not all) stages of reprogramming and help users understand both the strengths and limitations of the method.

We appreciate the reviewer's question regarding the modeling of cell state changes. Taking the example of in silico iPSC reprogramming, the model compares the rank value encoding of the original fibroblast transcriptome to the same rank value encoding but with OSKM added to the front of the rank value encoding. To evaluate the contextual awareness of the model, we then compare the resulting embeddings of the genes within the context of the original transcriptome to their embeddings within the context of the original transcriptome + OSKM. Each gene is affected within the self-attention computation by the remainder of the genes present in the encoded transcriptome, so the vast majority of the gene context remains equivalent to the original transcriptome aside from the addition of OSKM. Specifically, all of the fibroblast genes that are inactivated in later stages of the reprogramming are still present within the context presented to the model, and all of the iPSC genes that are activated downstream of OSKM are not yet present within the context presented to the model. Therefore, necessarily as a virtue of the mathematical computation that occurs in the model, the embedding of each gene is still largely reflecting the context of the fibroblast when considered in the absolute sense. This context is very different than what would be presented to the model if that gene was present in

the context of an iPSC, which has a very different set of genes expressed in its transcriptome compared to the original fibroblast.

However, the in silico reprogramming is modeling the *direction* of the shift in the embedding in response to adding OSKM. By comparing the original transcriptome to the original transcriptome + OSKM, we are defining the vector by which the genes' embeddings are shifted in response to this change in context. It is mathematically necessarily only modeling the first step in the reprogramming; essentially the model is being presented with a transcriptome that is the state of the cell immediately after OSKM are transfected/transcribed before they have had any effect on their downstream targets to shift the cell towards an iPSC. Based on this however, Geneformer is predicting the direction of the shift that will occur due to the addition of OSKM, which in the embedding space is statistically significantly towards the embedding position of the genes within the full iPSC context as compared to in silico reprogramming with random genes.

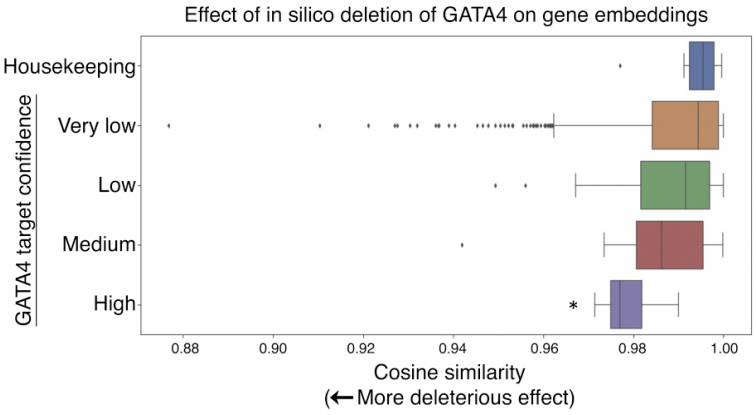
Predicting the cascade of cell state transitions that occurs after this first step of adding the reprogramming factors is of great interest to our group, and we are actively pursuing modeling approaches to accomplish this by applying similar techniques as are used in language modeling for next sentence prediction. As an analogy, given the initial sentence in a paragraph (fibroblast) and the first word of the next sentence (OSKM), we would like to generate the following sentences until the end of the paragraph (iPSC) to fully model the cell state transition from fibroblast to iPSC. However, this is a significant endeavor that we believe extends beyond the scope of this work.

To address the reviewer's point and ensure that the in silico reprogramming modeling approach is clear to users, we added the following discussion to the methods:

"Of note, in silico reprogramming/differentiation is only modeling the first step in the cell state transitions that occur in response to addition of the reprogramming/differentiation factors as the remainder of the gene context within the rank value encoding is equivalent to the original cell aside from the addition of these factors (e.g. OSKM or MYOD). The relevant measurement that is quantified in this approach is the directionality of the shift that occurs in response to addition of these factors. For example, while addition of random genes shifts the remaining genes' embeddings in a random direction, addition of OSKM significantly shifts the remaining genes' embeddings towards their embeddings within the context of the iPSC state, indicating that the directionality of the predicted shift is concordant with the true biological cell state transition from fibroblast to iPSC in response to the experimental addition of OSKM."

Lastly, we suggest that the authors compare the gene perturbations predicted by Geneformer in response to in silico deletion of TBX5 or GATA4 to gene expression changes in a RNA sequencing dataset of TBX5 or GATA4 knockout. For instance, in GATA4 knockout some direct targets are predicted to change more than others. We believe comparing changes predicted by Geneformer to a RNA sequencing dataset of the knockout will validate whether Geneformer is faithfully modelling gene knockout.

We appreciate the reviewer's point; indeed the most significantly differentially expressed genes in response to GATA4 knockout by experimental RNA-seq are more significantly impacted by in silico deletion of GATA4 in Geneformer modeling compared to those that are less significantly differentially expressed (as well as compared to housekeeping genes). We present this data in Extended Data Fig. 9d:



Referee #3 References:

1. Nelakanti,R. V and Xiao,A.Z. (2020) A New Link to Primate Heart Development. Dev. Cell, 54, 685–686.

2. Pirruccello,J.P., Bick,A., Wang,M., Chaffin,M., Friedman,S., Yao,J., Guo,X., Venkatesh,B.A., Taylor,K.D., Post,W.S., et al. (2020) Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat. Commun., 11, 2254.

3. Luna-Zurita,L., Stirnimann,C.U., Glatt,S., Kaynak,B.L., Thomas,S., Baudin,F., Samee,M.A.H., He,D., Small,E.M., Mileikovsky,M., et al. (2016) Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. Cell, 164, 999–1014.

4. Zhou,Y., Liu,Z., Welch,J.D., Gao,X., Wang,L., Garbutt,T., Keepers,B., Ma,H., Prins,J.F., Shen,W., et al. (2019) Single-Cell Transcriptomic Analyses of Cell Fate Transitions during Human Cardiac Reprogramming. Cell Stem Cell, 25, 149-164.e9.

Reviewer Reports on the Second Revision:

Referee expertise:

Referee #1: systems bio

Referee #2: network analysis

Referee #3: cardiac/computational biology

Referees' comments:

Referee #1 (Remarks to the Author):

Regarding the normalization process. The process is described in the main manuscript (line 109): "Each single cell's transcriptome is then presented to the model as a novel rank value encoding where genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M".

As known, if the normalization process of the training samples involves the test samples, this is usually considered as a wrong practice that may encode information about the test data in the training data.

I would like to ask the authors to verify whether the cells used as training samples have been normalized separately from the cells used as test samples. If yes, as I expect is the case, I would suggest mentioning it in the paper.

Referee #3 (Remarks to the Author):

We thank the authors for addressing the comments provided during the last revision. The authors have addressed our concerns where they have deemed it possible and included alterations in the text where the changes were not feasible. The use of genome wide association study phenotypes has provided an additional link with human physiology and development through complex trait biology. The potential of gene former to interpret GWAS results to improve understanding of complex trait biology should be emphasised further in the text.

The main revision required at this stage is in panel F of figure 6. The authors have failed to include a wild-type control in the panel to show the magnitude of improvement in contractile function after knockout of predicted genes. Including a wild-type control is crucial to understand to what degree the predicted genes rescue the dilated cardiomyopathy phenotype.

Author Rebuttals to Second Revision:

The reviewers' comments are shown in **black** and our responses are in **blue.**

Referee One:

Regarding the normalization process. The process is described in the main manuscript (line 109):
"Each single cell's transcriptome is then presented to the model as a novel rank value encoding where genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M".

As known, if the normalization process of the training samples involves the test samples, this is usually considered as a wrong practice that may encode information about the test data in the training data.

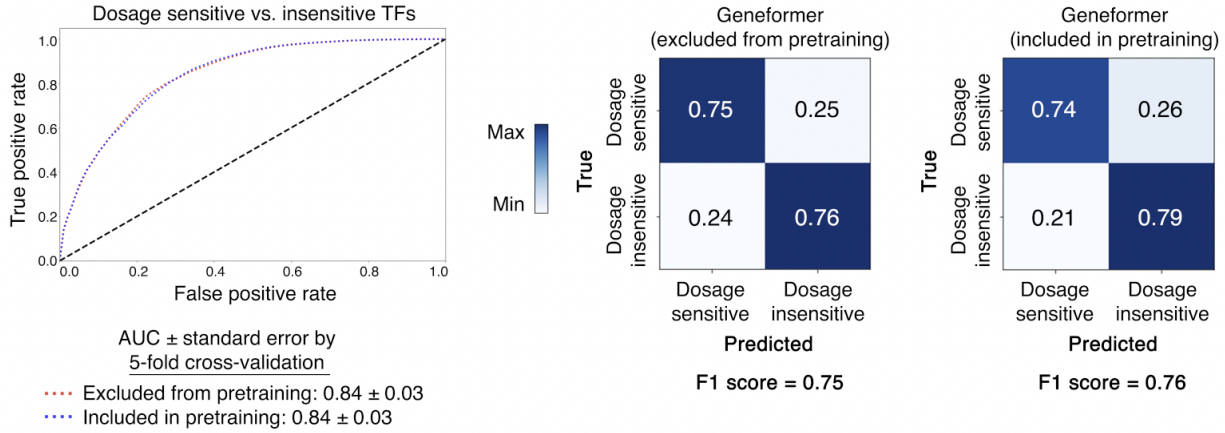
I would like to ask the authors to verify whether the cells used as training samples have been normalized separately from the cells used as test samples. If yes, as I expect is the case, I would suggest mentioning it in the paper.

Authors' response: We thank the reviewer for this comment. The model was pretrained using the pretraining corpus Genecorpus-30M, and the normalization factor for each gene is a stable factor derived from the cells in the pretraining corpus. The purpose of this normalization is to take advantage of the many observations of each gene's expression across Genecorpus-30M to prioritize genes that distinguish cell state. When new datasets are tested or used for fine-tuning, the normalization factor for each gene is not re-calculated based on the genes' expression in these cells. They use the same stable factor for each gene that was initially calculated. For example, the new datasets used for the cell type classification and disease modeling applications were normalized using the same stable factors for each gene as was used for pretraining.

We also would like to note that the same normalization process was used to generate the rank value encodings used to train the alternative modeling approaches (logistic regression, random forest, SVM, non-pretrained attention-based models), such that this was controlled for when comparing the efficacy of these alternatives.

In Extended Fig. 1f, we also address more broadly the possibility that there could be overlap between the large-scale corpus used for pretraining and smaller task-specific datasets utilized by future users for fine-tuning applications. This is analogous to large-scale pretrained models in the natural language understanding field where the large amount of text used for pretraining

could potentially overlap with text utilized by users for fine-tuning applications. The fine-tuning applications are trained on classification objectives that are completely separate from the masked learning objective used for pretraining, that although highly beneficial for pretraining is not practically useful as a downstream task. Therefore, the downstream tasks are a completely separate learning domain than predicting the identity of masked genes within a cell, and the inclusion of the data in pretraining is not relevant to the downstream application. To demonstrate this, we repeated pretraining on a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for the downstream application. We then fine-tuned the pretrained model to distinguish dosage sensitive vs. insensitive transcription factors using either the 10,000 held-out cells or 10,000 cells included in the subsampled pretraining corpus. The performance was equivalent, as shown in Extended Data Fig. 1f and below.



Manuscript changes:

To further ensure the normalization process is clear to users, we expanded the description with the following statements in the methods:

“This normalization factor for each gene is calculated once from the pretraining corpus and is used for all future datasets presented to the model. The provided tokenizer code includes this normalization procedure and should be used for tokenizing new datasets presented to Geneformer to ensure consistency of the normalization factor used for each gene.”

Referee Three:

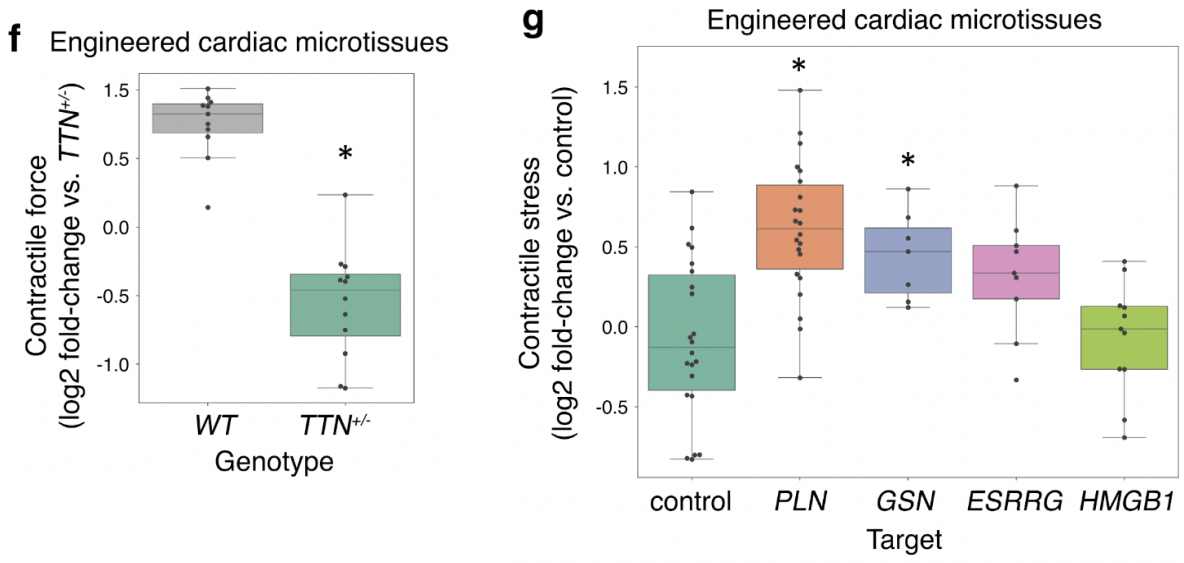
We thank the authors for addressing the comments provided during the last revision. The authors have addressed our concerns where they have deemed it possible and included alterations in the text where the changes were not feasible. The use of genome wide association study phenotypes has provided an additional link with human physiology and development through complex trait biology. The potential of gene former to interpret GWAS results to improve understanding of complex trait biology should be emphasized further in the text.

Manuscript changes: We thank the reviewer for this suggestion. We emphasized this point further within the discussion:

"Geneformer's ability to predict dosage-sensitive disease genes through the context-aware in silico deletion approach represents a valuable asset for interpretation of genetic variants, including prioritization of GWAS hits driving complex traits, and the specific tissues they are expected to affect."

The main revision required at this stage is in panel F of figure 6. The authors have failed to include a wild-type control in the panel to show the magnitude of improvement in contractile function after knockout of predicted genes. Including a wild-type control is crucial to understand to what degree the predicted genes rescue the dilated cardiomyopathy phenotype.

Manuscript changes: We thank the reviewer for raising this point. We had previously included the wild-type data in the supplementary materials but now moved it to the main Fig. 6 so that it can be interpreted alongside the gene targeting data.



Reviewer Reports on the Third Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have adequately addressed all my inquiries and I have no further comments. I congratulate the authors for this very impressive paper.

Referee #3 (Remarks to the Author):

The authors have addressed my concerns